A Genomics Perspective on RNA

5

Juliana C. Olliff, Jia A. Mei, Kristie M. Shirley, and Sara J. Hanson

Contents

J. C. Olliff
Department of Molecular Biology, Colorado College, Colorado Springs, CO, USA
e-mail: j_olliff@coloradocollege.edu

J. A. Mei · K. M. Shirley · S. J. Hanson (✉)
Department of Molecular Biology, Colorado College, Colorado Springs, CO, USA
e-mail: j_mei@coloradocollege.edu; k_shirley@coloradocollege.edu; shanson@coloradocollege.edu

N. Grover (ed.), *Fundamentals of RNA Structure and Function*, Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-90214-8_5

Keywords

mRNA-seq · Transcriptome · High-throughput sequencing · Genomics · Sequencing-by-synthesis · Differential expression · Long-read sequencing

What You Will Learn

- Introduction to sequencing technologies that allow for whole genome exploration of RNA, and how they have shaped the scale of RNA studies.
- Overview of the general workflow for a genome-scale experiment with RNA, including library preparation, sequencing, and analysis.
- Applications for sequencing methods to study a variety of RNA types, including messenger RNA and long or small noncoding RNA.
- Applications for sequencing methods to study processes involving RNA, including translation, transcription initiation, and RNA–protein interactions.
- Discussion of recent technologies that are expanding in the field, including long read, direct RNA sequencing, and single-cell sequencing.

5.1 From Transcript to Transcriptome: The Impact of Next-Generation Sequencing Methods on the Study of RNA

In recent decades, the development of tools for generating nucleic acid sequences at a high-throughput scale revolutionized the field of molecular biology. At the turn of the twenty-first century, methods for studying RNA were largely limited to the examination of individual genes (Fig. 5.1). The changes in expression of a single gene, for example, could be studied through Northern blots [1] and reverse transcriptase polymerase chain reaction (RT-PCR) [2]. Although powerful methods that allow for detailed investigation of a particular transcript, their reach is limited to the study of one or a few genes at a time. Scaling these studies to examine every gene in the genome is cost prohibitive and time

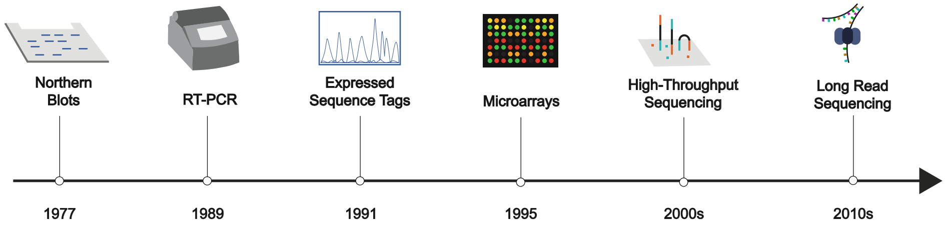


Fig. 5.1 Timeline of methods for studying gene expression

consuming. The first methods for studying RNA at a larger scale included cataloguing of Expressed Sequence Tags (a type of sequencing library produced from complementary DNA, or cDNA) by Sanger sequencing methods [3] and the development of microarrays [4]. Through these methods, the first glimpses of the genomic view of RNA were gleaned, but they remained limited by their cost, speed, and the resources required for their implementation.

With the advent of high-throughput sequencing (HTS), the volume of sequencing data that could be produced began to rise exponentially as the cost of sequencing dropped [5]. This led to a dramatic expansion in the scale of RNA studies, in which the perspective can be broadened to the transcriptome, or the entirety of a cell's or tissue's RNA content. The accessibility and affordability of these methods have led to them becoming standardized in the field, and have spurred the further development of cutting-edge technologies to delve into the complexities of RNA expression and function in the cell with much higher resolution [6].

HTS includes a set of methods and technologies that allow for simultaneous sequencing of millions of short (~50–300 bp), uniformly sized nucleic acid fragments. This chapter serves as an introduction to how HTS is used to study RNA at the transcriptome level. An overview of the workflow commonly used for high-throughput sequencing of messenger RNA (mRNA-seq) is presented, as well as modifications to the procedure that allow for the study of different types of RNA and cellular processes that involve RNA. Finally, more recent innovations in sequencing technologies and their applications are presented.

5.2 Using mRNA-Seq to Investigate the Protein-Coding Transcriptome

Given its role as the intermediate between DNA and protein, messenger RNA (mRNA) has been the focus of large amount of transcriptome work. mRNA sequencing (mRNA-seq) has been used to address a wide array of biological questions. How does gene expression change in response to environmental conditions, or during different stages of development? When and where are different isoforms of an RNA transcript expressed? Which genes are regulated by a particular transcription factor?

Sequencing-by-synthesis (SBS), developed by Illumina, Inc., is a popular HTS method for performing mRNA-seq. Like other HTS methods, SBS technology allows for massively parallel sequencing of short nucleic acid molecules. Depending on the specific sequencing platform used, an SBS run can produce as much as six trillion base pairs (6 Tb) of sequence in a single run [7]. The volume of data produced by SBS allows for accurate and robust quantification of gene expression across the transcriptome [8], which can be used for comparison of gene expression levels between samples. SBS can also allow for paired-end sequencing, in which both ends of a DNA molecule are sequenced. The additional sequencing information can be used to assemble transcriptomes de novo, or without the need for a preexisting genome sequence for an organism to be used as a reference [9]. In addition, SBS protocols can be modified to track the strandedness, or directionality, of a transcript, which distinguishes between sense and antisense transcription [10]. Careful planning of a sequencing experiment includes considering how the final dataset will be analyzed in order to prepare a sequencing library and sequencing reaction that will sufficiently address the biological question of interest [11].

This section delves into the workflow and methods commonly used to perform mRNA-seq with SBS (Fig. 5.2). RNA is first isolated from a sample and mRNA is used to generate a cDNA library. Following sequencing of the cDNA library, bioinformatics methods are used to assemble the short sequences into transcripts, and further analysis is performed, such as identifying transcripts that are differentially expressed between experimental conditions.

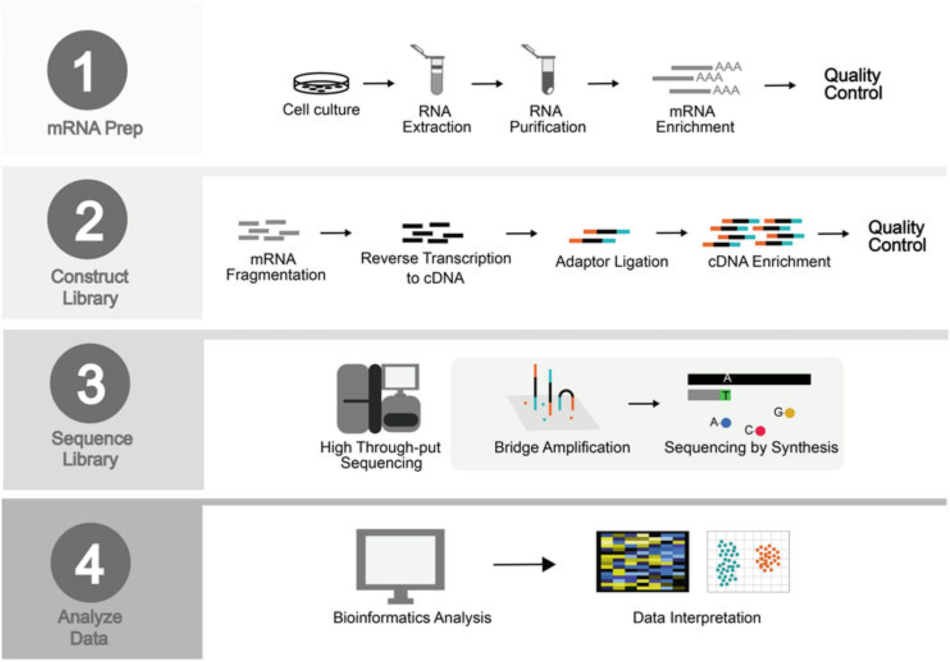


Fig. 5.2 Overview of mRNA-seq workflow. An mRNA-seq experiment using SBS technology occurs in a series of four stages. (1) mRNA is isolated from a sample or samples of interest. (2) A cDNA library is constructed from the input mRNA. Library construction includes addition of adaptor sequences that are required for the downstream sequencing reaction. (3) The cDNA library is amplified on a solid substrate and SBS determines the sequence of each fragment in the library in parallel. (4) The large amount of sequencing data is processed and analyzed to address biological questions

5.2.1 Preparing RNA for Sequencing: Isolation, Fragmentation, and Enrichment

The first step in performing an mRNA-seq workflow is isolation of RNA from a source of interest, such as a multicellular tissue or a population of unicellular organisms. This step must be performed with biological replication by preparing multiple samples that have received identical treatment (typically a minimum of three for each sample type in the experiment) to allow for statistical evaluation during data analysis. To isolate RNA, cells must be chemically lysed using detergents to disrupt cell membranes. In the case of cells with cell walls, such as plants and yeasts, mechanical disruption is also required using glass beads or mortars.

Following lysis, RNA must be isolated from other macromolecules found in the cell, including DNA, proteins, and lipids. The differing solubilities of these macromolecules are used as the basis for extraction. Proteins and lipids are removed first due to their solubility in organic solvents (phenol and chloroform, respectively). In contrast, the nucleic acids RNA and DNA are soluble in aqueous solutions, allowing for their separation from the organic solutes. When isolating RNA from a sample, acid phenol chloroform is used. The reduced pH in this method encourages the solubility of DNA in the organic phase, thereby enriching the aqueous phase specifically with RNA [12]. To further purify the RNA from the nucleic acid mixture, samples are treated with deoxyribonuclease I (DNase I), an enzyme that breaks down double-stranded and single-stranded DNA molecules without sequence specificity [13].

Working with RNA requires a great deal of care to prevent the degradation of the RNA molecules. Unlike DNA, RNA contains a hydroxyl group on the $2'$ carbon of each nucleotide that is susceptible to hydrolysis reactions that will break down the backbone of an RNA molecule. In addition, ribonucleases (RNases), enzymes that target RNA molecules for degradation, are ubiquitous in intracellular and extracellular environments. Using RNase inhibitors can help mitigate the degradation of samples, in addition to using careful lab practices and sterile technique.

The RNA purity, quality, and quantity are rigorously assessed prior to constructing a sequencing library. RNA sample purity is measured to determine the amount of genomic DNA contamination that is present in the sample following DNase treatment. The RNA will be converted to cDNA during library construction, and the presence of genomic DNA may result in incorporation of genome sequences into your library that do not reflect levels of gene expression. Low quantities or poor-quality RNA can also impact cDNA library prep. Library preparation requires established minimum amounts of RNA, and samples with quantities less than the minimum may result in libraries that do not represent the data accurately [14]. Degraded RNA samples can also have a substantial impact on the accuracy of expression quantification during transcriptome analysis [15]. Therefore, RNA with minimal degradation and high levels of integrity are preferred when possible. In some cases, the sample type can make isolation of high-quality RNA a challenge. For example, biopsy specimens that are formaldehyde fixed, paraffin-embedded (FFPE) are subject to crosslinking of RNA with other macromolecules which then dramatically decreases RNA yield and quality [16]. When this is the case, the analysis of these datasets should be adjusted to take the RNA quality into account [15].

Because quantity and quality of RNA are so critical to the accuracy of a transcriptome, multiple complementary methods are used for their measurement. There are three main techniques for measuring RNA quality: (1) ultraviolet (UV) spectroscopy, (2) fluorometry, and (3) size separation.

UV spectroscopy can be used to quantify and differentiate between types of macromolecules because of their characteristic absorbance wavelengths [17]. While this can easily distinguish between some types of macromolecules, such as nucleic acids and proteins, it can be much more difficult to distinguish between DNA and RNA, which absorb similar wavelengths. UV spectroscopy is therefore useful to determine the amount of protein contamination in your sample but will not provide a measurement of DNA contamination.

Fluorometry can provide a more accurate measurement of RNA quantity in a nucleic acid sample by differentiating between its DNA and RNA composition. In this method, the samples are treated with fluorescent dyes that bind specifically to the molecule of interest (e.g., double-stranded DNA, single-stranded DNA, or RNA). Measurement of the fluorescence of the sample is determined by a fluorometer and this value is used to calculate the concentration of the specific nucleic acid type in a sample.

Size separation is the standard method used to assess RNA sample quality. In a sample of total eukaryotic cellular RNA, 80% of molecules are ribosomal RNA (rRNA) [18]. In eukaryotes, this includes the 28S rRNA molecules that are part of the large ribosomal subunit, and the 18S rRNA molecules that comprise the small ribosomal subunit (23S and 16S in bacteria). If the 28S and 18S rRNA molecules are intact, the overall integrity of all RNA molecules in the sample can be inferred to also be of high quality. To examine this, a small portion of the RNA sample is run through a gel matrix in a capillary tube to separate all of the RNAs by size, and the size distribution of the sample is measured. In a high-quality RNA sample, this will yield two large peaks that correspond to the 18S and 28S rRNAs (Fig. 5.3). If the ratio of the quantity of 28S to 18S rRNA is at least 2.0, the sample is considered to be of high quality. This ratio is used to calculate a value known as the RNA integrity number (RIN), which ranges between 1.0 (RNA is completely degraded) and 10.0 (RNA is completely intact) [19].

Once a sample has been determined to have sufficient quality to use for sequencing, the RNA sample is enriched for the RNAs that are of interest in the study. Because rRNA makes up the vast majority of the total RNA content in cells, preparation of an unenriched RNA sample will result in nearly all of the sequencing data representing these molecules. Several methods are available for enrichment of a sample of RNA, including selection for polyadenylated (polyA) RNA species, depletion of rRNA, or cDNA capture.

For mRNA-seq, a commonly used enrichment method is polyA selection. mRNAs (in both eukaryotic and bacterial cells) are polyadenylated: during transcription, a string of adenine nucleotides is added on to the 3′ end of the RNA molecule [20] (Chap. 6,

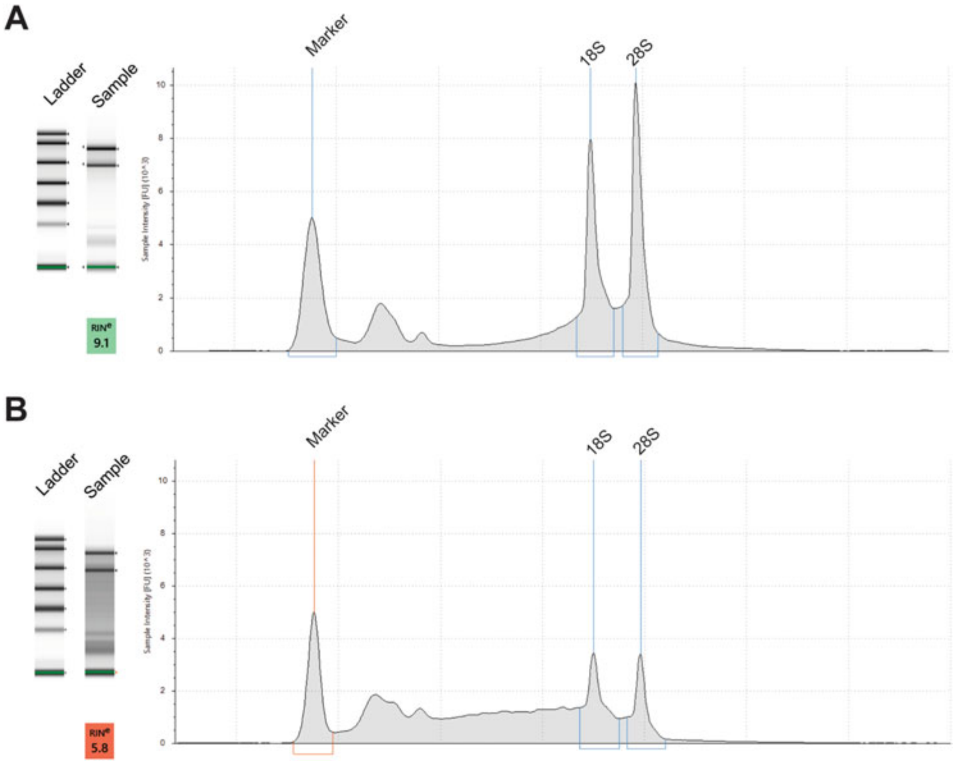


Fig. 5.3 Using size selection to assess RNA sample quality. Example traces showing the size distribution for RNA samples with (**a**) high integrity and (**b**) low integrity. Images of the size separation of the sample compared to ladder are shown on the left, and densitometry traces indicating the amount of fluorescence signal detected across the size range are shown on the right. The samples were analyzed using an Agilent 4150 TapeStation platform

mRNA). In this method, the 3′ polyA tails found on mature mRNAs are targeted for selection using magnetic beads coated with polyT oligonucleotide probes. These probes form complementary base pairs with the polyA tails of the mRNA molecules, allowing them to be pulled out of solution on a magnetic rack. RNAs that do not contain a polyA tail, such as rRNA, will be washed away and removed from the sample, quickly paring down the type of RNA to be sequenced. An alternative enrichment method specifically targeting rRNA molecules is ribosomal depletion, which is described in more detail in the section on sequencing long noncoding RNAs below.

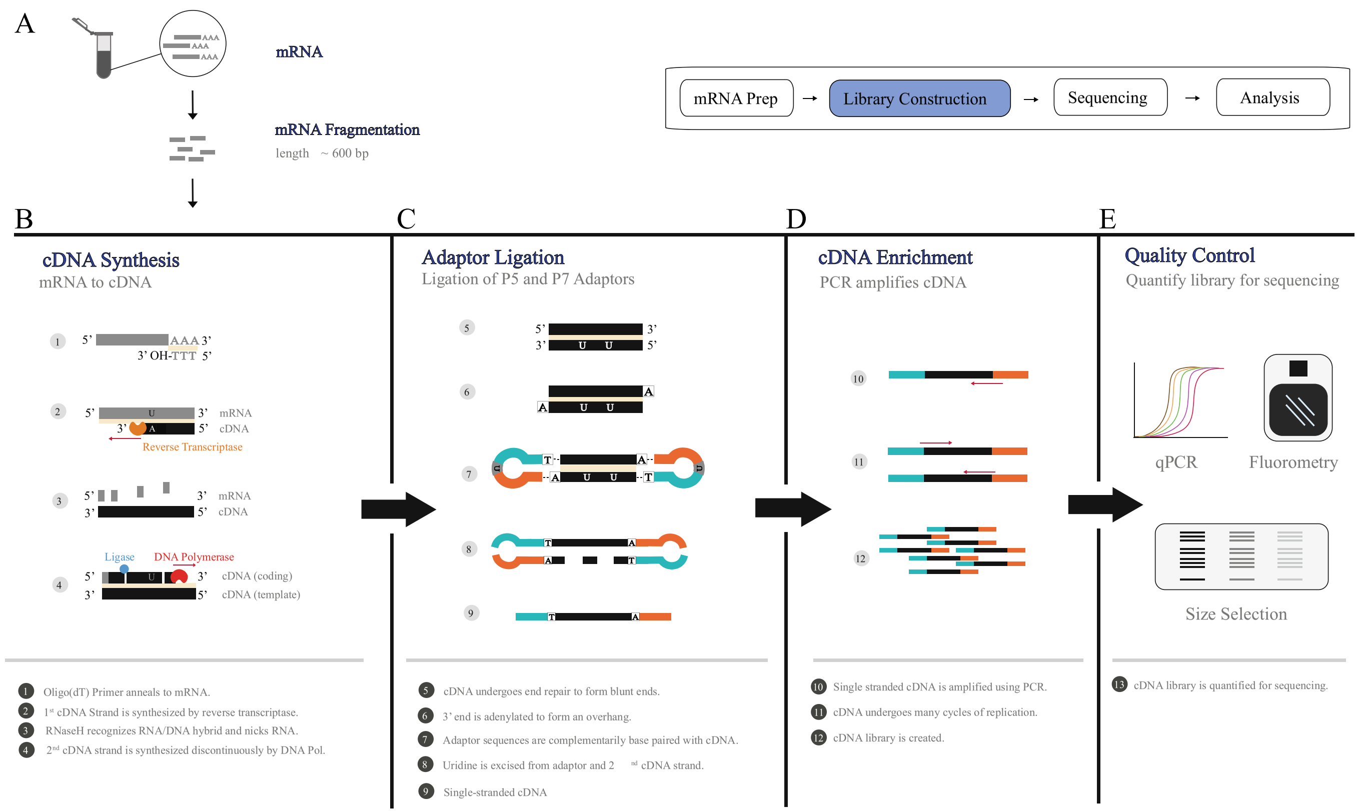
Although polyA enrichment is an efficient way to enrich for mRNAs, it introduces notable biases into a sample, particularly the exclusion of any transcript that is not polyadenylated, which includes some types of long noncoding RNAs [21]. Additionally, because the mRNAs are selected for their 3′ end, any degradation or fragmentation of transcripts in the sample will cause the 3′ end of the transcript to be overrepresented relative to the rest of the molecule in the final sequencing library. This makes the integrity of the input RNA sample critical to the success of the final sequencing library.

Enrichment for an mRNA-seq experiment can also be performed using cDNA capture. This method is highly specific and can only be used when the sequences of interest in the sampled organism are known in advance. Rather than using polyT oligonucleotide probes, magnetic beads are coated with customized sequence that will complement the target sequences. This method is typically performed after the RNA sample has been used for cDNA synthesis (see below). The cDNA is allowed to base pair with the probes, while any molecules that do not match are washed away. This method enriches for specific transcripts to be sequenced, making it useful for focusing on a small set of known genes of interest. For example, cDNA capture has been used in the study of gene expression in different types of human cancer cells. Many mutations in coding sequences that contribute to cancer have been previously identified, and cDNA capture has been used to identify expression levels for the mutated alleles in cancerous cell types [22]. cDNA capture is a useful enrichment method for this type of experiment because it allows for targeting specific sequences, removing all background noise and homing in on a small number of genes of interest. In addition, cDNA capture can be useful when the genes of interest are expressed at low levels. By enriching the samples for these genes, they can be more readily identified and quantified in the sample [23].

5.2.2 Constructing Sequencing Libraries: Strandedness, Multiplexing, and Amplification

Once a high-quality RNA sample has been enriched for the RNA species of interest, the next step is to construct a sequencing library. Library construction is a multi-step process (Fig. 5.4) in which the RNA is fragmented and used as a template for cDNA synthesis. The cDNA molecules are then modified to contain adaptor sequences that are required for the sequencing reaction.

Although great care is taken during RNA isolation to ensure that samples are not degraded, the first step of mRNA-seq library preparation is to fragment the mRNA molecules to ~600 bp pieces (Fig. 5.4a) using physical, chemical, or enzymatic methods [24, 25]. This step is important for the quality of the sequencing library, as the final sequences generated during the sequencing reaction will be of a much shorter length than most mRNA molecules ($\leq$300 bp). In addition, the length of molecules in the library can introduce bias at several points during library construction, including during cDNA synthesis and PCR enrichment. By fragmenting the RNA into a uniform size distribution, these biases can be reduced [10]. Another advantage provided by fragmentation is the reduction of secondary structures that would inhibit efficient library construction. These structures, like hairpins or clovers, form through intramolecular base-pairing and are less likely to form in shorter RNA molecules.



1 Oligo(dT) Primer anneals to mRNA.
2 1st cDNA Strand is synthesized by reverse transcriptase.
3 RNaseH recognizes RNA/DNA hybrid and nicks RNA.
4 2nd cDNA strand is synthesized discontinuously by DNA Pol.

5 cDNA undergoes end repair to form blunt ends.
6 3' end is adenylated to form an overhang.
7 Adaptor sequences are complementarily base paired with cDNA.
8 Uridine is excised from adaptor and 2 nd cDNA strand.
9 Single-stranded cDNA

10 Single stranded cDNA is amplified using PCR.
11 cDNA undergoes many cycles of replication.
12 cDNA library is created.

13 cDNA library is quantified for sequencing.

Fig. 5.4 Creating a stranded mRNA-seq library for SBS**.** Preparing an mRNA-seq library for SBS requires (**a**) mRNA fragmentation, (**b**) cDNA synthesis with dUTP-labeling of the coding strand, (**c**) sequencing adaptor ligation and removal of the second synthesized cDNA strand, (**d**) amplification of the cDNA library by PCR, and (**e**) quantification of the cDNA library by qPCR, fluorometry, and size selection

After fragmentation, RNA is used as a template to produce cDNA. cDNA synthesis provides an advantage because of the increased stability of a DNA molecule relative to RNA, allowing for easier handling and storage of the sample. In addition, SBS technology relies on DNA-based methods like Polymerase Chain Reaction (PCR—Box 5.1). Creation of a cDNA library from the RNA sample is therefore required for the downstream sequencing process.

Box 5.1 Polymerase chain reaction

Polymerase Chain Reaction (PCR) is a foundational method in molecular biology to amplify a targeted sequence of DNA (Fig. 5.5). DNA replication is carried out in vitro by combining a template DNA sequence, short oligonucleotide primer sequences that are synthetically produced to target a sequence of interest, free nucleotides, and a thermostable DNA polymerase. The reaction is subjected to 25–35 cycles of three steps. (1) The reaction is heated to a high temperature (95 °C) to *denature* the template DNA, or break the hydrogen bonds between base pairs. The DNA polymerase used in this reaction was isolated from the thermophilic bacteria *Thermus aquaticus* and can withstand this high temperature without itself being denatured. (2) The temperature is reduced (45–65 °C) to allow base pairs to form between the template DNA and the primer sequences. (3) The temperature is increased (72 °C) to provide optimal conditions for DNA synthesis. During this step, DNA polymerase uses the free 3′ hydroxyl of the primer to initiate synthesis of a new DNA strand that is complementary to the target DNA sequence. With each repetition of this cycle, the number of target sequences in the sample is doubled, resulting in 2^n copies of the target sequence, where n equals the number of cycles performed.

A critical consideration during cDNA library construction is whether your sample will be stranded or unstranded. When creating a stranded library, the information regarding the strand of DNA that was transcribed is maintained. During transcription, the DNA strand whose sequence is identical to the transcript is known as the coding strand, while the DNA strand whose sequence was used for complementary base-pairing by RNA polymerase is known as the template strand (Chap. 10, Transcription?). Stranded library preparations can therefore differentiate between transcription that took place to create a sense strand of RNA that matches the coding strand for a known gene, or transcription that took place to create an antisense strand of RNA that matches the template strand of a gene. Antisense transcription has been shown to play important roles in transcriptional regulation [26], making preservation of the strandedness of a library critical to the investigation of these functions. The increased amount of information provided by stranded libraries has led to them becoming standardized in the field for mRNA-seq library preparation protocols [10, 27].

To make a stranded cDNA library (Fig. 5.4b), a pool of random short oligonucleotides are added to the RNA samples, which can anneal to the RNA and act as primers by providing an available hydroxyl group on the 3′ carbon of the last nucleotide of the

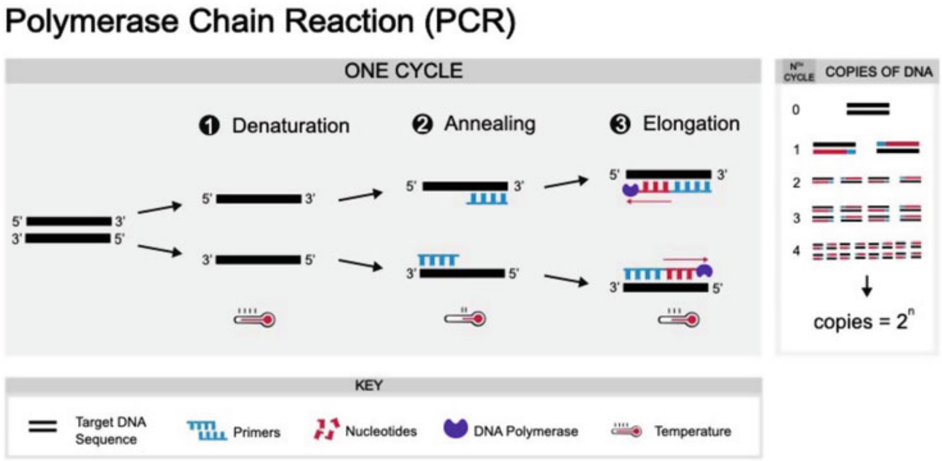


Fig. 5.5 Polymerase chain reaction (PCR). The events of one cycle of PCR are shown (denaturation, annealing, elongation). Amplification of a targeted DNA sequence by PCR is exponential, with the final number of copies equal to approximately 2^n, where n = the number of PCR cycles

oligonucleotide. The enzyme reverse transcriptase uses this free 3′-OH for incorporation of nucleotides to synthesize a new cDNA strand from the RNA template in the 5′ to 3′ direction. This first cDNA strand corresponds to the template DNA strand that is complementary to the transcript sequence. After reverse transcription, the first strand of cDNA remains bound by complementary base-pairing to the RNA template, creating an RNA/cDNA hybrid. The enzyme RNaseH, which recognizes RNA/DNA hybrid molecules, is then used to create nicks in the RNA molecule. DNA polymerase I from the bacteria *Escherichia coli* uses the broken RNA molecule as primers to synthesize a second strand of cDNA, complementary to the first that reflects the sequence of the coding DNA strand. The second strand created through this method is discontinuous. *E. coli* DNA ligase is used next to create covalent phosphodiester bonds that complete sugar-phosphate backbone of the second strand of cDNA. During second-strand synthesis, DNA polymerase I is provided with a pool of nucleotides that includes deoxyuridine triphosphate (dUTP) rather than deoxythymidine triphosphate (dTTP). This marks the second strand of cDNA as distinct from the first, which will provide the strandedness of the library.

Next, the double-stranded cDNA molecules are modified to facilitate their use in a sequencing reaction (Fig. 5.4c). The cDNA first undergoes end repair using T4 DNA polymerase, *E. coli* DNA polymerase I, and T4 polynucleotide kinase to generate molecules with blunt and phosphorylated 5′ ends before adding single 3′-A ends of the molecules. This overhang allows for complementary base-pairing with a 5′-T overhang found on adaptor sequences that are added to the ends of each cDNA fragment. The adaptors contain known DNA sequences that will be used during the sequencing process and are added to the cDNA molecules as hairpin loops including a uracil nucleotide in the center. The hairpin loop structure increases ligation efficiency and decreases adaptor dimerization events, as there is less steric hinderance during ligation.

Following adaptor ligation, the uracil nucleotides are excised from the second strand of cDNA and from the hairpin loops of the adaptors using uracil DNA glycosylase to remove the uracil base and either enzymatic or chemical cleavage to digest the backbone at the abasic site. This linearizes the adaptor sequences and creates gaps in the second strand of cDNA, resulting in a single intact cDNA strand (the first synthesized strand), which corresponds to the complement of the original RNA molecule (the template DNA strand). Alternatively, some methods for stranded mRNA-seq library construction use a polymerase in the PCR enrichment step below that cannot synthesize DNA past a dUTP nucleotide. The second strand of cDNA is excluded as a PCR template in this method because it cannot be amplified. In either method, the end result is a clear differentiation between the cDNA strands to identify the template and coding strands of DNA for each mRNA.

The single-stranded cDNA library remaining is then enriched using PCR (Fig. 5.4d). Extension primers that complement the adaptor sequences are used to amplify the cDNA. These primers contain additional sequences required in the SBS sequencing reaction, known as P5 and P7 sequences. The PCR enrichment results in cDNA library fragments with a P7 sequence on the 5′ end of the cDNA strand that corresponds to the coding strand of DNA, and a P5 sequence on the 5′ end of the cDNA strand that corresponds to the template strand of DNA. This provides the directionality required to determine the strandedness of each sequence generated in the sequencing reaction. In addition, the P7 sequence also contains a barcode that is unique to a sample, allowing for multiplexing of samples during the sequencing reaction. In a multiplexed sequencing reaction, multiple samples are pooled and sequenced simultaneously with the sample barcodes allowing the sequences from each sample to be separated following the sequencing reaction. Once PCR enrichment is complete, the library is carefully quantified using quantitative PCR (qPCR), fluorometry, and size selection (Fig. 5.4e) and is loaded onto the sequencing platform.

5.2.3 Generating the Transcriptome: Bridge Amplification and Sequencing-by-Synthesis

Sequencing of a cDNA library for mRNA-seq using SBS technology [28] occurs in two parts (Fig. 5.6). First, each cDNA molecule in the library is amplified on a solid-state flow cell, resulting in clusters of identical sequences that will serve to increase the sequencing signal. Second, cycles of reversibly terminated nucleotide incorporation occur. This allows for fluorescently labeled nucleotides to be added one at a time, with each incorporated nucleotide detected as the strand is synthesized. After 50–300 cycles of nucleotide incorporation, this process is often repeated for sequencing from the other end of the cDNA fragment to produce paired-end data.

SBS occurs on a flow cell, which is a glass slide containing multiple channels, also referred to as lanes. During the sequencing reaction, reagents needed for the sequencing

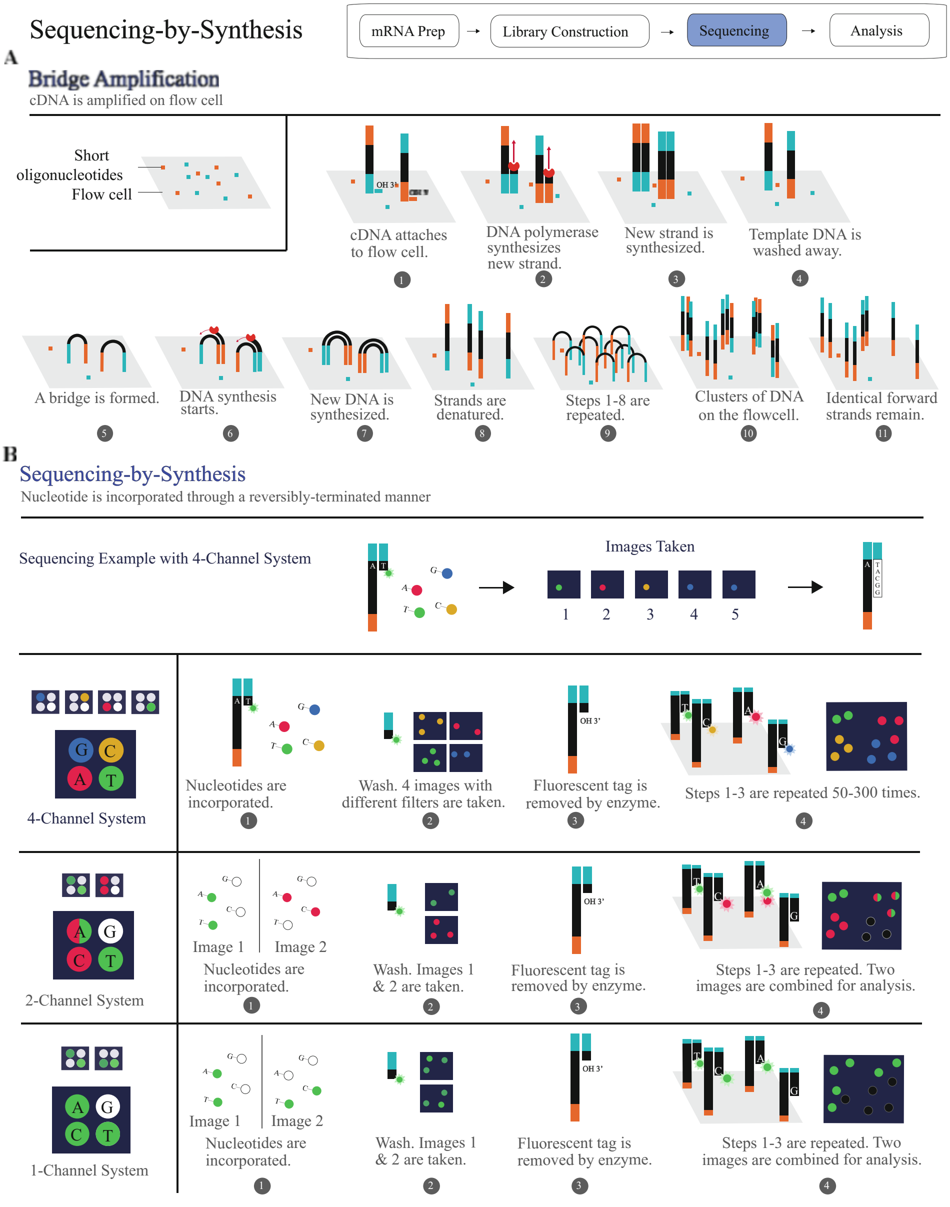

Sequencing-by-Synthesis
mRNA Prep
Library Construction
Sequencing
Analysis
A
Bridge Amplification
cDNA is amplified on flow cell
Short oligonucleotides
Flow cell
OH 3'
cDNA attaches to flow cell.
1
DNA polymerase synthesizes new strand.
2
New strand is synthesized.
3
Template DNA is washed away.
4
A bridge is formed.
5
DNA synthesis starts.
6
New DNA is synthesized.
7
Strands are denatured.
8
Steps 1-8 are repeated.
9
Clusters of DNA on the flowcell.
10
Identical forward strands remain.
11
B
Sequencing-by-Synthesis
Nucleotide is incorporated through a reversibly-terminated manner
Sequencing Example with 4-Channel System
A T
G
A
T
C
Images Taken
1 2 3 4 5
A T A C G G
G C A T
4-Channel System
Nucleotides are incorporated.
1
Wash. 4 images with different filters are taken.
2
OH 3'
Fluorescent tag is removed by enzyme.
3
T C A G
Steps 1-3 are repeated 50-300 times.
4
A G C T
2-Channel System
Image 1
Image 2
Nucleotides are incorporated.
1
Wash. Images 1 & 2 are taken.
2
Fluorescent tag is removed by enzyme.
3
Steps 1-3 are repeated. Two images are combined for analysis.
4
A G C T
1-Channel System
Image 1
Image 2
Nucleotides are incorporated.
1
Wash. Images 1 & 2 are taken.
2
Fluorescent tag is removed by enzyme.
3
Steps 1-3 are repeated. Two images are combined for analysis.
4

Fig. 5.6 Sequencing-by-synthesis (SBS)**.** SBS begins with (**a**) bridge amplification to create clusters of identical library fragments on a flow cell. After cluster generation, (**b**) sequencing-by-synthesis begins with cycles of reversibly terminated nucleotide incorporation and base-calling that occurs using 4-channel, 2-channel, or 1-channel chemistry

chemistry are introduced into the lanes using microfluidics. The first step of sequencing is to attach the cDNA library fragments to the flow cell and generate identical copies that will form a cluster for each library fragment. The flow cell surface is coated with covalently attached short oligonucleotides that are complementary to the P5 and P7 sequences found on each cDNA fragment (see above). The cDNA library is denatured and allowed to hybridize with these oligos, and the oligos can then serve as primers for PCR to generate sequence clusters through a process called bridge amplification (Fig. 5.6a).

During bridge amplification, base-pairing between the flow cell oligo and the cDNA library fragment provides a free 3′ hydroxyl that allows DNA polymerase to begin synthesis. DNA polymerase continues synthesis to create a full complementary strand. The double-stranded molecule is denatured, and the template cDNA molecule is washed away, leaving the newly synthesized strand that was extended from the covalently attached oligo. The new strand folds over to form a "bridge" and the non-anchored end hybridizes to another complementary oligonucleotide bound to the flow cell surface. A new cDNA template strand is amplified, forming a double-stranded cDNA molecule. The strands are denatured, and the process repeats several times. Finally, the reverse strands (corresponding to the template DNA strands) are cleaved from the oligo sequence and washed away, leaving a cluster of identical forward strand molecules (corresponding to the coding DNA strands). The cluster undergoes the same sequencing reaction in the next step, and the simultaneous incorporation of identical fluorescent nucleotides amplifies the sequencing signal. Each fragment of cDNA forms its own cluster, leaving the flow cell coated with millions (or billions, depending on the sequencing platform) of clusters that are now ready for the sequencing reaction.

The sequencing reaction (Fig. 5.6b) is initiated through the addition of primers that are complementary to the adaptor sequences incorporated during library construction. DNA polymerase uses the free 3′-hydroxyl provided by the primer to incorporate additional nucleotides. The free nucleotides provided to DNA polymerase for addition to the synthesized DNA strand contain two critical modifications. First, each type of nucleotide is tagged with a fluorescent dye (dTTP = green, dCTP = yellow, dATP = red, and dGTP = blue). Second, the 3′ carbon is attached to a reversible terminator molecule rather than a hydroxyl group. This modification ensures that only a single nucleotide can be incorporated by DNA polymerase [29].

SBS occurs as the reaction cycles through a series of steps. (1) DNA polymerase and modified free nucleotides are added to the reaction, and DNA polymerase incorporates the next complementary nucleotide. The unincorporated nucleotides are washed away, and (2) the fluorescent color emitted by the cluster is recorded. (3) Enzymes are then added to the reaction to remove the fluorescent tag and the terminator molecule from the 3′ carbon, providing a free hydroxyl that will allow for extension by DNA polymerase in the next cycle. These steps are repeated 50–300 times, and the order of fluorescent color emissions for each cluster on the flow cell is interpreted as a DNA sequence.

Notably, some sequencing platforms that use SBS (such as Illumina NextSeq and Illumina iSeq) have adjusted technology to use a two color (2-channel) or one color

(1-channel) system, respectively, rather than 4-channel to increase the efficiency of the sequencing run (Fig. 5.6b). In a 4-Channel system, a different color is used to identify each type of nucleotide, each requiring a separate filter for detection of its color, as described above. In 2-Channel sequencing, dTTP fluoresces green, dCTP fluoresces red, dATP fluoresces both green and red, and dGTP lacks fluorescence [30]. This combination requires only two filters, green and red, to detect all four nucleotides. In contrast, 1-channel sequencing methods label dTTP with green fluorescence, dATP with green fluorescence that can be enzymatically removed, dCTP with a linker group, and dGTP without fluorescence. Following nucleotide incorporation, an image is taken. The dATP fluorescence is then removed while fluorescent labels are added to the linker group attached to dCTP and a second image is taken. The pattern of loss or maintenance of the fluorescence state is used to infer the nucleotide that was incorporated during that cycle [31]. Using 1-channel SBS still requires two images during each cycle of sequencing, but needs only a single filter. Decreasing the filter number required for sequencing reduces the processing of the colors during each cycle of the sequencing run. Although the 1- and 2-channel systems are more prone to errors due to the increased probability of misinterpreting a fluorescent signal, the overall error rate for sequencing using this method remains very low.

Following the 50–300 cycles of nucleotide incorporation and fluorescence detection, a collection of single-end sequences have been collected, with one sequencing read (the raw sequence produced during the sequencing reaction) generated per cDNA library fragment. However, the sequencing process can be repeated to produce a second read for each fragment, beginning from the opposite end of the cDNA fragment, and sequencing in the opposite direction. To accomplish this, once the first read has been generated, bridge amplification is repeated and the forward strand is cleaved and washed away from the flow cell. New sequencing primers are added and the reverse strand is used as a template for SBS. Samples that undergo this additional round of sequencing now have paired-end reads, with each cDNA molecule having two sequencing reads that reflect each end of the fragment [32].

Paired-end sequencing provides several advantages over single-end data. Paired-end reads provide not only twice the amount of sequencing data, but also allow for spatial inferences. The cDNA fragments that make up the sequencing library are of known approximate length due to the RNA fragmentation step during sequencing library preparation. Therefore, the length of the unknown sequence that separates the two reads can be used to aid in assembly of transcripts during analysis (see below). Whether or not this additional information is required for mRNA-seq depends on the nature of any particular experiment, the availability of genome or transcriptome sequences for the organism of interest, and the planned data analysis.

5.2.4 Making Sense of the Data: Transcriptome Assembly and Differential Expression

Analyzing the very large amount of sequencing data generated in an mRNA-seq experiment progresses through a well-established workflow requiring specialized software (Fig. 5.7). The steps of mRNA-seq analysis typically include quality control and read processing, transcriptome assembly, and differential expression analysis [11, 33]. Once these steps are completed, a full picture of the genome-wide transcriptional changes in protein-coding sequences will be created.

The first steps of mRNA-seq analysis are focused on assessing the quality of the sequencing data and performing any necessary editing of the sequencing reads (Fig. 5.7a) [14]. The raw sequencing data is a large text file that contains both the nucleotide sequence of each individual sequencing read, and a quality score for every base call made by the platform during the sequencing reaction. The quality score reflects the statistical confidence in the base call for each individual nucleotide in every sequencing read that was generated. To assess the quality of the overall sequencing run, the attributes of the sequencing reads are analyzed, including examining the quality scores of the reads, the nucleotide composition (such as the GC content), the lengths of the sequence, and the frequency of identical reads in the sample. Filtering of the reads may be required at this point if any major issues are observed with the sample to remove low-quality sequences before additional analysis is performed. Reads are further processed by identifying and trimming any adaptor sequences before performing further analysis.

The next portion of mRNA-seq analysis is to use the short sequences produced by the sequencing reaction to assemble full-length transcripts (Fig. 5.7b). Depending on the experiment and the resources available for the organism being studied, the approach for this step can fit into one of two methods: reference-based or de novo assembly [34].

Reference-based assembly requires a previously generated genome or transcriptome assembly. For this method, the short sequencing reads are “mapped” to a reference sequence (genome or transcriptome). Mapping entails aligning the reads to the reference where the sequence is an identical match. After mapping, transcripts are assembled from overlapping reads. Challenges can arise for reference-based transcriptome assembly in organisms that contain repetitive genomic regions. In these cases, reads can map to multiple locations in the reference sequence, which makes the true transcriptional location ambiguous. An additional challenge occurs in organisms that contain introns and use alternative splicing. In this case, multiple transcript isoforms may exist for a single gene. Mapping software identifies splicing events using reads that span the junction between exons and therefore map to regions that flank intron sequences.

De novo assembly is required when a reference sequence is not available or is not of sufficient quality for the organism under study, which is most likely to be the case in studies of non-model organisms. De novo (“from nothing”) assemblies build transcripts with no prior knowledge of the transcriptome’s contents. Since no reference sequence is used, de novo assembly programs use de Bruijn graphs, which are a way of mapping paths through a

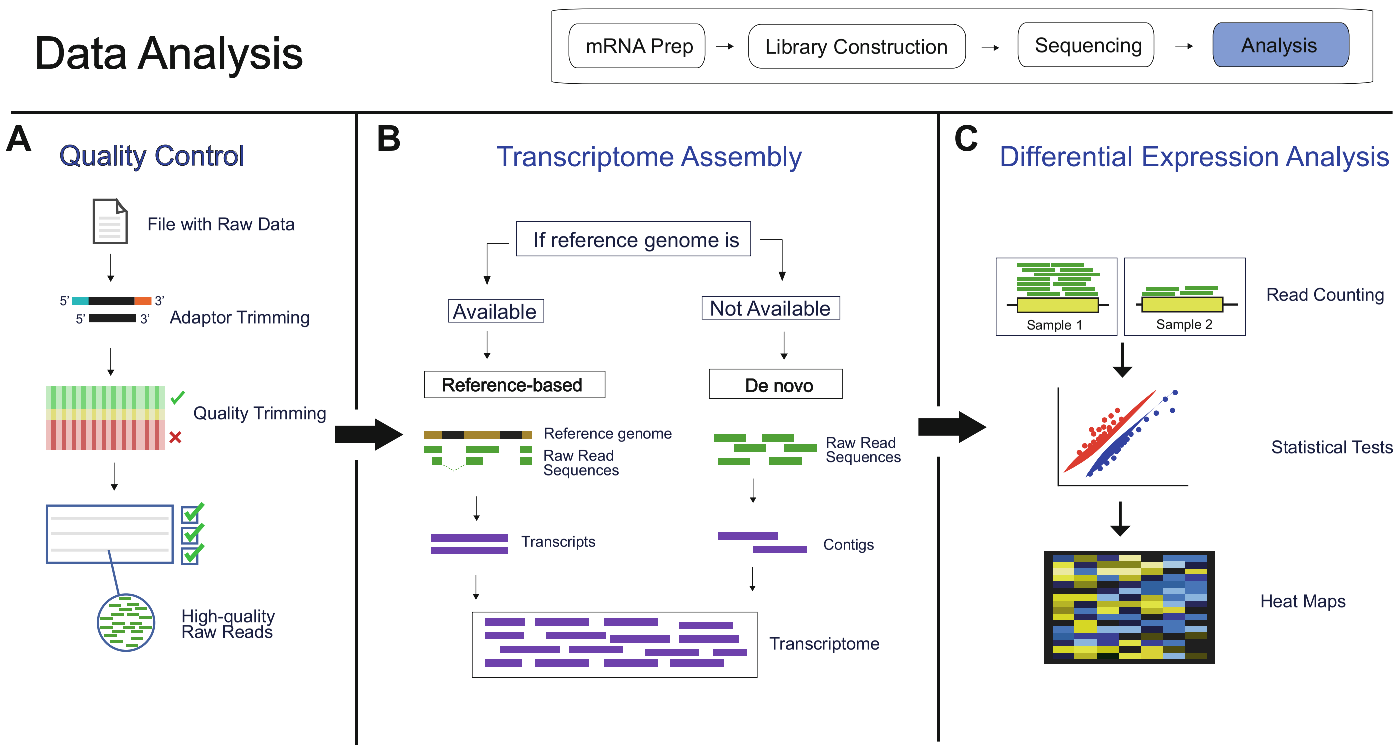


Fig. 5.7 mRNA-seq data analysis workflow. Analysis of mRNA-seq data requires (**a**) quality control steps to remove sequencing adaptors and low-quality sequencing reads and (**b**) assembly of the short reads into full-length transcripts using a reference (reference-based) or from scratch (de novo). In many experiments, (**c**) differential expression analysis is performed to quantify and compare levels of gene expression between two or more samples

sequence assembly to predict likely transcripts based on short overlaps in the read sequences. Paired-end data greatly facilitates this de novo transcriptome assembly because of the spatial information provided by having two reads from the same cDNA fragment that is of known approximate length. The final result of a de novo assembly is a set of contiguous sequences (contigs) that were built from the sequencing reads. If successful, there will be one contig representing each transcript. Since the only sequences present in an mRNA-seq dataset are polyA transcripts expressed under the experimental conditions, this assembly represents the set of coding sequences expressed under the experimental conditions rather than the entirety of the organism's genome sequence.

Whether reference-based or de novo assembly is used to build a transcriptome, an important consideration is the transcriptome variability that will occur between samples. A well-designed mRNA-seq experiment will always include biological replicates for a particular experimental condition, and will frequently include replicates generated under different experimental conditions. Although the biological replicates receive identical treatment between samples, biological and technical variability will result in differences in the assembled transcriptome between replicates. In addition, changes in experimental conditions may result in specific transcripts or isoforms only being represented in one condition or another due to specificity of their expression conditions. To ensure that the final transcriptome assembly is inclusive of all transcripts, the reads from all samples can be pooled to generate one single transcriptome that can then be used as a reference for read mapping or annotation [11]. Alternatively, software is available that will merge the transcript assemblies from multiple samples into one single holistic assembly [35].

For many mRNA-seq experiments, the next step of the analysis will require examining differential expression, or the changes in expression that occurred as a result of the experimental conditions (Fig. 5.7c). A powerful feature of mRNA-seq experiments is that the sequences produced not only tell you which genes are expressed, but also provide quantitative information regarding the level of expression occurring for every gene. The amount of expression is inferred by the number of reads that originate from a gene in a sample. This is determined through read mapping and counting the reads that map to a particular gene or transcript. For reference-based assembly, this information may already be captured through the assembly process. De novo assemblies will require an additional step, in which the reads are mapped to the newly generated transcriptome assembly.

Once the reads are mapped and counted, the quantity of reads for each gene or transcript is compared between the samples from each experimental condition. Statistical tests are performed to assess the likelihood each observed change in expression is greater than would be expected by random chance. These tests take into account the variation in expression observed between biological replicates, as well as the level of expression of a particular gene or transcript.

Using mRNA-seq to study differential gene expression has been a popular use of this technology as it provides insight into how the cell modulates transcriptional activity in a variety of cellular contexts. In addition, a wide range of experiments can also be designed using differential expression to learn about gene regulation and transcriptional networks.

By making controlled changes to the genome, such as gene deletions or insertions, and comparing mRNA expression in these cells to wild-type cells, the effects of these changes can be examined genome-wide. Experimental manipulation of the genome is particularly useful when it comes to understanding the effects of transcription factors on the expression of other genes.

The complexity of the transcriptome beyond differential expression can also be investigated using mRNA-seq methods [33]. The high resolution of transcript sequence information provided through mRNA-seq can allow for determination of allele-specific expression at heterozygous loci in the genome. This information can also be integrated with variation in gene expression levels across samples to identify expression quantitative trait loci (eQTL), which are sites that provide genetic contributions to this variation.

5.3 Beyond mRNA-Seq: Other High-Throughput Sequencing Applications for RNA

Focusing an experiment on RNA molecules that encode for protein products can be a powerful way to gain insights into cellular responses and pathways. However, other applications have also been developed that focus on other pieces of the transcriptome, such as long and small noncoding RNAs, or nascent transcripts. In addition, library construction has been adapted to study the interactions between RNA molecules and other cellular components, such as the ribosome and RNA-binding proteins. These tools expand the repertoire of sequencing-based methods that provide a picture of RNA function.

5.3.1 Long Noncoding RNA

Long noncoding (lncRNAs) are RNA transcripts greater than 200 base pairs long that perform a function as RNA molecules rather than encode for proteins. LncRNAs perform a wide range of important cellular functions, including tethering transcriptional machinery to DNA, increasing mRNA stability, and regulating transcription of other genes [21]. Preparing a sequencing library that includes lncRNAs is very similar to the library preparation procedure for mRNA-seq described, but can include some important modifications.

One critical difference between preparing an mRNA-seq library, and a total RNA-seq library that includes lncRNAs is in the enrichment method. In mRNA-seq, a sample of RNA was enriched for mRNAs that have a polyA tail. Although some lncRNAs are polyadenylated, many are not and would therefore be lost during library preparation using this method. Therefore, an alternative enrichment method known as rRNA depletion is often used instead. rRNA depletion is a more inclusive enrichment method that selectively *removes* one class of RNA (rRNA) from the sample rather than selectively *retain* one particular class of RNA molecule. During the enrichment step, the samples are mixed with magnetic beads coated with oligonucleotides that complement rRNA sequences found in the organism of interest. Once the rRNAs have hybridized with the oligos on the beads, the beads are pulled down on a magnet and discarded while the rest of the sample is kept for library preparation. This leaves behind a diverse pool of transcripts for sequencing, regardless of polyadenylation status.

Although an effective and commonly used enrichment method, there are some important considerations when using rRNA depletion for sample enrichment. Kits that have been manufactured for this method are typically made for rRNA sequences found in model organisms. Although these sequences are well-conserved, they may not effectively remove all rRNA molecules in non-model organisms when nucleotide differences in these sequences exist between species. In addition, cross-reactivity between sequences may occur if a sequence found in the rRNA probe matches a sequence elsewhere in the transcriptome. This would result in inadvertent removal of the transcript from the sample and its absence from the final sequencing data. To reduce these concerns, rRNA depletion can be performed through selective degradation rather than with magnetic beads. For selective degradation, a pool of longer complementary DNA oligos are added to the sample that will anneal to the entire length of the targeted rRNA. The RNA:DNA hybrids are then targeted for enzymatic degradation by RNaseH and DNaseI [36]. This method may be more effective for enrichment of a sample from a non-model organism whose RNA may not be removed entirely by a commercially available depletion kit. Additionally, this method can be adapted to target not only rRNAs, but also other highly expressed transcripts that perform constitutive functions in the cell (i.e., transcripts for housekeeping genes). This selective removal allows transcripts expressed at very low levels, such as many lncRNAs to be more easily discovered in the sample [21].

Following the enrichment step, the rest of the sequencing library construction occurs in the same way as for mRNA-seq. In the case of sequencing lncRNA, creating a stranded library is absolutely essential due to the antisense transcription used to produce many of these molecules [37]. Transcriptome assembly is also very similar to mRNA-seq. However, after transcriptome assembly is complete, candidate lncRNAs require further computational analysis of their coding potential and/or experimental evaluation of their structure and function [38].

5.3.2 Small Noncoding RNAs (miRNAs, tRNAs, etc.)

Small noncoding RNA (sncRNA) includes transcripts that do not encode for proteins and have a shorter length (often less than 100 nucleotides). This includes RNAs that function in regulating the expression other genes, such as microRNA (miRNA) and small interfering RNA (siRNA), and RNAs that are involved in maintaining genome integrity, such as piwi-interacting RNA (piRNA) (Chap. 8, ncRNA) [39]. RNA-seq methods have facilitated sncRNA discovery and quantified their expression under a variety of conditions, providing insights into their roles in critical cellular processes and in the development and progression of human disease [40].

Performing a sequencing experiment targeting sncRNAs differs from mRNA-seq in the enrichment stage of the library preparation protocol. In the case of sncRNAs, the enrichment methods focus on selecting for RNA species based on their size. Because coding mRNA molecules tend to be longer than 500 nucleotides, selection is performed for RNA molecules that are shorter than this length. This can be done by separating the RNA sample by size on a polyacrylamide gel, excising the desired size range of sample, and extracting the RNA. Methods and specialized equipment have been created to automate this process, which makes it much more precise and increases the RNA yield from the extraction. Following the enrichment, no fragmentation step is required, as the transcripts are already at a length that is short enough for the sequencing reaction.

In some cases, an alternative method of enrichment is used for sncRNA library preparation. Because many sncRNAs require interactions with specific RNA-binding proteins to carry out their functions, these RNA-protein interactions can be used to enrich for specific types of sncRNA. For example, miRNA bind to proteins in the Argonaute family to target specific mRNA molecules for degradation or inhibit their translation in numerous eukaryotic organisms [41]. To enrich for miRNAs in a sample, antibodies that recognize the argonaute proteins can be used for immunoprecipitation that will pull down both the argonaute proteins and any RNA molecules that they interact with. Although this method is much more specialized than size selection, and would not be useful for examining the entire small noncoding transciptome, it allows for focus on a particular subset of sncRNAs and may allow for discovery of molecules expressed at low levels that would be obscured by other enrichment techniques.

Generating a cDNA sequencing library for sncRNAs also differs from mRNA-seq library preparation methods. For sncRNA library preparation, adaptors are ligated to the 3′ end of the RNA molecules first [42]. Primers are then added that hybridize to the 3′ adaptor before ligating adaptors to the 5′ end of the RNA molecules. The sample is then enriched for RNA molecules that contain the primer and both adaptors before cDNA synthesis and PCR enrichment is performed. These modifications increase the yield of molecules represented in the final sequencing library. However, a challenge of sncRNA library preparation methods is the bias that can be introduced. Molecules that are expressed at low levels may not have accurate quantitative representation in the final library. In addition, sncRNAs in some organisms contain posttranscriptional modifications that make adaptor ligation less efficient. Adjustments to the adaptor sequences and reaction conditions can help mitigate these biases for some types of sncRNAs [42].

Analysis of a sncRNA sequencing library is very similar to the mRNA-seq workflow. Following quality control, read mapping, and differential expression analysis, additional work may be required to further investigate the functional implications of the data. For example, miRNA sequences of interest may be analyzed to predict their mRNA target sequences. Conclusions from these types of analyses require further investigation and validation through experimental work.

5.3.3 Investigating RNA Biology: Other Applications of RNA Sequencing

In addition to studying gene expression dynamics, sequencing methods can be used for large scale examinations of cellular mechanisms that involve RNA. These types of methods focus on a variety of aspects of RNA biology, including RNA structure, transcription, translation, and other mechanisms that involve RNA–protein interactions. A brief sampling of these methods is provided below, although many more continue to be developed to pursue a variety of questions related to RNA.

Studying Translation Using Ribo-Seq Although mRNA-seq methods will provide information about the transcripts present under the conditions of a particular experiment, it does not allow direct inferences about the proteins being synthesized. Regulation of the timing and rate of translation are critical processes in the control of gene expression by the cell. To examine translation dynamics, all mRNA molecules that are interacting with ribosomes are sequenced (Ribo-Seq) providing a snapshot of all mRNAs being actively translated [43]. To accomplish this, cells are treated with the antibiotic cycloheximide to cause translation to stall. The treated cells are then lysed, RNA is extracted, and RNases are used to enzymatically degrade all RNA that is not bound to the ribosome. The mRNA-ribosome complexes are purified using centrifugation or chromatography, and mRNA is purified from the ribosomes for use in creating cDNA sequencing libraries. The sequences generated provide a ribosome profile that shows not only *which* mRNA molecules were being translated, but *which part* of a particular mRNA was being translated. It also provides quantitative information that can be used to infer rates of translation for each mRNA. Ribo-seq data can be used to identify novel open reading frames (ORFs), alternative translation initiation sites in known ORFs, and can be used in combination with mRNA-seq to examine transcript stability and post-transcriptional gene regulation.

Studying Transcription Through Nascent RNAs The mRNA-seq methods described above reflect the steady state of RNA found in the cell, or the summation of RNA that is produced and degraded in a particular condition. However, RNAs that are inherently unstable, such as RNAs that are produced at enhancer sequences (enhancer RNAs, or eRNAs) that have functions in regulating enhancer activity, are not easily examined this way. In addition, the dynamics of the process of transcription are not reflected in mRNA-seq datasets. Insights into RNA polymerase pausing, transcriptional termination, and RNA modifications that occur concurrently with transcription (capping, splicing, and polyadenylation) (Chap. 6, mRNA; Chap. 10, transcription) can be investigated by examining nascent RNAs, which are those molecules that are in the process of being transcribed [44].

Several approaches can be taken to examine nascent transcription using high-throughput sequencing methods [6, 44]. (1) RNA that is associated with chromatin can be isolated using salt washes (caRNA-seq). These samples would include not only nascent RNAs, but also spliceosomal RNAs and any lncRNAs that functionally associate with DNA. (2) RNA polymerase II, which is responsible for transcription of mRNAs, small nuclear RNAs, and miRNAs, can be tagged with a small epitope. Following isolation of chromatin, the epitope tag can be used to immunoprecipitate RNA polymerase II and any RNAs with which it is associated. The RNAs are then purified from RNA polymerase II and used to create a sequencing library (mNET-seq). (3) Run on methods can be used, in which transcription in the cell is halted through drug treatment or freezing temperatures, nuclei are isolated, and transcription is allowed to resume in vitro with nucleotide analogs such as 5-bromouridine 5′-triphosphate (BrUTP). BrUTP-containing RNAs, which represent those that are newly transcribed, can then be enriched in the final sample by targeting the analogs for immunopurification (GRO-seq). (4) In a manner similar to a run on method, newly produced RNAs can be examined in vivo by providing cells with labeled nucleotides in their media and allowing them to be incorporated for a defined period of time. Following RNA isolation, the samples can then be enriched for these nascent transcripts by targeting the labeled molecules through immunopurification (TT-seq, TimeLapse-seq, or SLAM-seq).

Studying RNA–RNA and RNA–Protein Interactions RNA function can depend on its interactions with other RNA molecules (e.g., miRNA binding to the 3′ UTR of target mRNAs) or with RNA-binding proteins. A variety of methods have been developed to examine these interactions using high-throughput sequencing technologies to define a cellular “interactome” [6]. RNA–RNA interactions can be identified by using biotinylated psoarlen, which intercalates into RNA–RNA hybrid molecules and crosslinks the interactions (*S*equencing of *P*soralen crosslinked, *L*igated, *A*nd *S*elected *H*ybrids, or SPLASH). The crosslinked molecules can be purified using streptavidin to pull down the biotinylated molecules. Following purification, the RNA hybrids are fragmented and ligated to join the interacting molecules into a single strand that is used for library preparation and sequencing. Intermolecular RNA-RNA interactions are inferred during analysis based on the sequencing reads that contain more than one RNA species.

To examine interactions between RNA and proteins, immunoprecipitation with antibodies targeting a protein of interest is performed either without crosslinking or following ultraviolet (UV) crosslinking of the cells (RIP-seq and CLIP-seq, respectively). UV crosslinking will form covalent bonds between RNA and protein, but will not crosslink protein–protein interactions, which reduces the noise in the sample. Following immunoprecipitation, the RNA is purified from the protein and used as the input for library preparation and sequencing. Analysis of these datasets can reveal all RNAs that interact with a particular protein and can be used to identify the sequence motifs found in the RNA molecule that are recognized by the protein.

Studying RNA Structure RNA molecules can form secondary structures through intramolecular base-pairing interactions, as well as tertiary structures. These structures are critical to their functions and interactions with other molecules in the cell (Chap. 2, Architecture of RNA) [45]. The "structurome" can be established through high-throughput sequencing methods [46]. To determine the unstructured, single-stranded regions and structured, double-stranded regions of RNA molecules, selective enzymatic digestion of single-stranded or double-stranded RNA is performed (PARS or FRAG-seq). The remaining RNA is used for library preparation and sequencing. Performing the experiments to create ssRNA and dsRNA sequencing libraries in parallel provides a full picture of the RNA structurome under the experimental conditions. Chemical mapping can be used as an alternative to nuclease treatment, in which chemical probes can be used to mark structured or unstructured RNA and targeted to enrich the final library prep (SHAPE-seq).

5.4 The Present and Future of Transcriptomics

Short-read sequencing-by-synthesis methods have served as a powerful means to generate quantitative surveys of gene expression for many types of RNA across a diversity of organisms. Now that well-established protocols are in place to conduct differential expression analysis and de novo transcriptome assembly, there is wide implementation of them in a range of experimental frameworks.

More recent developments have focused on expanding these transcriptome studies to incorporate new technologies that address some of the key limitations of short-read HTS sequencing applications to transcriptome work, as well as expand the technological toolbox for studying gene expression to include methods aimed at better understanding heterogeneity in gene expression and RNA modifications.

5.4.1 Improving Assembly: The Advent of Long-Read Sequencing Methods

Following the rise and success of short-read SBS methods developed by Illumina, a so-called "third-generation" or "next-next generation" of sequencing technologies has now emerged and expanded. Key innovations in these technology developments have included *increasing the length* of sequencing reads and *removal of pre-sequencing sample processing steps* to allow for more efficient and less biased data. In particular, the increased length of the sequencing reads allows for greatly improved de novo assembly of the transcriptome and enhances the detection of transcript isoforms with less ambiguity [47].

These new methods of sequencing differ from short-read SBS methods in two key ways: sequencing a *single-molecule* of DNA or RNA at a time and carrying out sequencing of a molecule *directly*. In particular, two methods have gained popularity, each having distinct sequencing mechanisms underlying their technology.

The first method of third-generation long-read sequencing is single-molecule real-time (SMRT) sequencing, developed by Pacific BioSciences (PacBio) [48]. To conduct transcriptome studies using this method, RNA is first converted into cDNA, and hairpin adpators are ligated to the ends of the cDNA molecules. This creates a circular single-stranded molecule that is used for sequencing. Following adaptor ligation, the cDNA library is loaded onto a SMRT cell, which is composed of small wells called zero-mode waveguides (ZMWs). Each ZMW contains a DNA polymerase enzyme fixed to the bottom. The DNA polymerase binds to the cDNA molecule and initiates DNA synthesis using the cDNA molecule as a template (Fig. 5.8a). During synthesis, the DNA polymerase is provided four differentially labeled fluorescent nucleotides that will emit a pulse of light once the nucleotide is incorporated into the synthesized strand. The series of pulses generated during strand synthesis are detectable by an imager and are interpreted as a DNA sequence.

There are two key features of SMRT sequencing that distinguish this method from the short-read SBS method implemented by Illumina [48]. First, sequences reflecting entire RNA molecules can be represented in the final dataset, as no RNA fragmentation is required during library preparation prior to cDNA synthesis. Second, no bridge

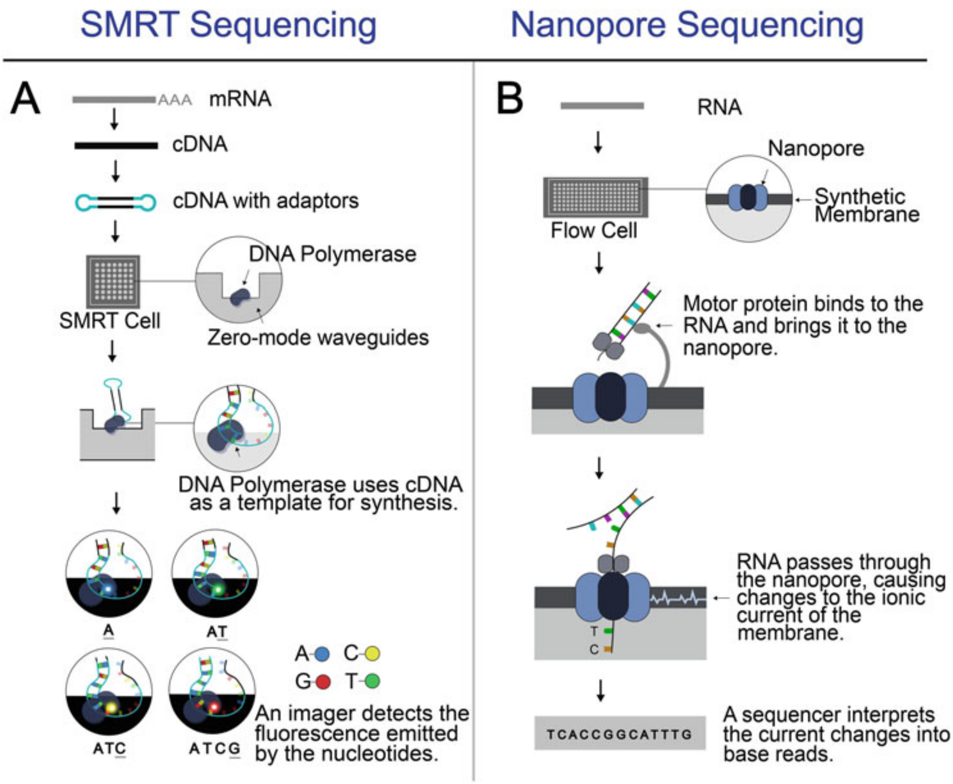

SMRT Sequencing
Nanopore Sequencing
A
AAA
mRNA
cDNA
cDNA with adaptors
SMRT Cell
DNA Polymerase
Zero-mode waveguides
DNA Polymerase uses cDNA as a template for synthesis.
A
AT
ATC
ATCG
A ● C ●
G ● T ●
An imager detects the fluorescence emitted by the nucleotides.
B
RNA
Nanopore
Flow Cell
Synthetic Membrane
Motor protein binds to the RNA and brings it to the nanopore.
RNA passes through the nanopore, causing changes to the ionic current of the membrane.
T
C
TCACCGGCATTTG
A sequencer interprets the current changes into base reads.

Fig. 5.8 Long-read sequencing methods. Long-read sequencing methods include (**a**) SMRT sequencing and (**b**) nanopore sequencing

amplification step occurs prior to SMRT sequencing; the fluorescent signal emitted by the single synthesized molecule in the ZMW is detected by the imager without the need for a cluster of identical molecules. Removal of the need for fragmentation and sequence amplification for SMRT sequencing reduces the errors and biases that can be introduced by these steps. In addition, the length of reads that are produced through SMRT sequencing is substantially increased at up to 25 kilobases per read. This increased length comes at a cost of total reads produced, however, as a typical run of SMRT sequencing yields 4,000,000 reads (compared to the up to 20 billion reads produced by the Illumina NovaSeq platform). In addition, the error rates in SMRT sequencing are higher than in SBS (~15% vs. ~0.1%, respectively).

The second method of third-generation sequencing marketed by Oxford Nanopore Technologies, high molecular-weight RNA molecules are sequenced directly without the need for cDNA conversion [49]. The sequencer contains a synthetic membrane with hundreds or thousands of nanopores, depending on the sequencing platform used (Fig. 5.8b). RNA molecules bound by motor proteins are brought to the nanopores. As the RNA is passed through the nanopore, it disrupts an ionic current that is formed by a preestablished voltage gradient across the membrane. The alterations to the ionic current as the molecule moves through the nanopore are detected by the sequencer, with the shape of each nucleotide creating a characteristic change to the current. These changes are used by the sequencer to infer base calls and generate a sequencing read.

The feature that distinguishes nanopore sequencing from both SBS and SMRT sequencing is that the bases are called *directly* from the RNA molecule. No cDNA synthesis or PCR amplification is required as part of the library preparation or sequencing reaction. This removes several forms of sequencing bias that have been observed through these other methods. In addition, detection of ionic current changes does not require an imager, which reduces the size of the sequencing equipment to as small as the palm of a hand. This makes the technology portable, so it can be brought into the field for immediate processing of clinical or environmental samples.

As with SMRT sequencing, nanopore sequencing results in much longer read lengths, lower total read output, and a higher error rate than SBS. Nanopore sequencing can yield up to 242 gigabases of sequencing data with read lengths limited only by the length of the RNA fragments in the sequencing library (reads in the megabase range length have been achieved). However, the error rate of the base-calling by this method is typically ~10%.

Long reads generated by third-generation sequencing methods are the basis for isoform-sequencing (Iso-seq). The length of reads sequenced by SMRT and nanopore methods greatly enhances the assembly of a transcriptome and the detection of splice isoforms in a sample. Because the length of the reads is on the order of several kilobases, which is well within the average length of an mRNA transcript in humans (3522 bp) [50], *Drosophila melanogaster* (3058 bp) [51], or yeast (~1250 bp) [52], a full-length transcript can be sequenced in its entirety as a single read and does not require computational assembly. This overcomes the challenges in transcriptome assembly from SBS data of identifying splice-junctions and in assembling alternatively spliced transcripts. Thus, with generation of a

SMRT or nanopore transcriptome, the library of transcript isoforms can be identified fully and without bias or ambiguity introduced through the computational methods of assembly [49].

In important limitation in long-read sequencing technologies is the increased error rate in the reads relative to short-read sequencing [47]. Although some forms of bias in sample generation are removed through these methods, others are introduced, and the reported error rate for PacBio SMRT (~15%) and Oxford Nanopore (~10%) sequencing methods is substantially higher than SBS (~0.1%). This can be addressed in the downstream analysis by increasing the *coverage* of the transcriptome (i.e., sequencing multiple reads per transcript), thus allowing for the representation of more correct than incorrect sequences and allowing a consensus sequence to be inferred. With SMRT sequencing specifically, this can also be mitigated through the generation of circular consensus sequences (CCS). CCS generation takes advantage of the hairpin adaptors/circular structure of the sequenced molecule. By continuing to synthesize DNA from the circular molecule, you can sequence the same molecule multiple times, generating a long read that can be chopped up and assembled into a consensus sequence that represents the fragment [48]. By generating redundancy in the sequencing reaction, you can identify individual nucleotide errors and remove them from your downstream analysis.

5.4.2 Examining the Epitranscriptome: Direct Detection of RNA Modifications

Posttranscriptional modifications to both noncoding and coding RNA molecules can impact their structure and function by influencing their stability, localization, and interactions with other molecules. Well-known modifications to RNA, such as *N*6-methyladenosine, 5-methylcytosine, 7-methylguanosine, pseudouridine, and adenine to inosine editing, can be detected and quantified through SBS methods [53]. However, using SBS results in indirect detection and requires extensive library preparation that often involves immunoprecipitation.

The development of long-read sequencing methods has led to the ability to sequence RNA molecules directly (Direct RNA-seq), which can allow for direct detection of posttranscriptional modifications. Nanopore sequencing has been used successfully to detect a variety of RNA modifications [54]. Just as the shape of the individual bases changes the ionic gradient across the membrane in specific ways, unique signatures are detected when a base is modified. Thus, as the molecule passes through the membrane, the bases can be read along with their modifications to generate not only the sequence of the RNA, but to also determine which bases have been modified, and with which particular modification.

SMRT sequencing has also been tailored to allow for the detection of RNA modifications [54]. By modifying the sequencing reaction to use reverse transcriptase instead of DNA polymerase for nucleotide incorporation, the RNA molecules can be sequenced directly without the need for cDNA library construction. The kinetics of nucleotide incorporation during SMRT sequencing are changed in consistent and predictable ways when the template contains modifications. The altered kinetics can be interpreted by the sequencer as particular posttranscriptional modifications. Although detection of nucleotide modifications using SMRT sequencing has more frequently been applied to genomic DNA sequencing, it has also been successfully applied to transcriptome sequencing.

5.4.3 Deciphering Heterogeneity: Transcriptomes from Individual Cells

The standard workflow for mRNA-seq library preparation that is described above is limited to examining populations of cells such as a pool of unicellular yeast growing the same condition, or a particular tissue from a mouse that is homogenized prior to extraction of RNA. The gene expression data that results from this type of library preparation therefore reflects an average across the cells that were used for extraction. Preserving heterogeneity in gene expression across cells in a dataset can provide valuable insights into important aspects of cell biology, such as cell type identification and function within a tissue, cellular differentiation, and drug resistance in cancer treatment [55]. In order to examine the variation in gene expression between individual cells in a population or tissue, methods for single-cell RNA sequencing (scRNA-seq) have been developed.

To perform scRNA-seq, individual cells must be separated prior to RNA isolation [6, 55]. A variety of methods have been used to isolate individual cells for scRNA-seq. These include diluting a sample to the level of a single cell and micromanipulation or microdissection to isolate individual cells from under a microscope. Other methods include fluorescence-activated cell sorting (FACS), in which fluorescently labeled cells are identified when they pass by a laser and are then separated from the rest of a population, and microfluidics to isolate individual cells in nanoliter-sized oil droplets.

Following single cell separation, RNA isolation and sequencing library preparation are performed. Sample processing for the individual cells that have been isolated occurs in steps very similar to the library preparation described previously, with two critical differences [6, 55]. First, each single-cell sample must be labeled with a unique barcode. These barcodes allow each sequencing read to be assigned to the cell from which it originated. Second, the amount of RNA yielded from a single cell is much smaller than in a typical bulk RNA-sequencing approach. First- and second-strand cDNA synthesis are performed on the RNA isolated from the individual cells (which can be done with polyA selection for analysis of mRNAs). However, to generate enough samples for sequencing, the cDNA is typically amplified via PCR. The amplification step can introduce substantial biases in the final sequencing library, which can be mitigated by the use of unique molecular identifiers (UMIs). UMIs are barcode sequences added to each library fragment during cDNA synthesis. After amplification, all library molecules that share a barcode can be traced back to a single starting RNA molecule, which allows for correction of any biases in the data during analysis.

This workflow has been further modified to allow for the preservation of spatial information for cells within a tissue [6, 55]. Isolation of a specific tissue section through laser capture microdissection (LCM) prior to single-cell separation can provide resolution on the location of cells with particular patterns of expression. Alternatively, mRNAs can be directly isolated from a tissue section by overlaying the tissue on a microarray chip with barcoded oligodT probes (Slide-seq). The barcodes are used to retain the spatial information for each RNA molecule in the tissue during data analysis.

The challenges in analyzing scRNA-seq data come first from associating the gene expression to a particular cell or cell type, then from filtering out the technical and biological variation that exists between samples and correcting for sample bias generated during library preparation [55]. Once these quality control steps are taken, the power of scRNA-seq allows for novel cell-subtypes to be defined within a sample, gene regulatory networks (genes that are coordinately regulated) to be elucidated, and cell fate specification to be determined.

Take Home Message

Exploration of RNA biology using genome-scale methods is now standard practice in molecular biology. The field of transcriptomics continues to expand and evolve, with new technologies and methods being developed at a rapid pace. Improvement to sequence quality and read length promise to continue to make these methods approachable for investigating a wide range of biological questions. Beyond the broad patterns of gene expression that can be readily assessed through sequencing, we can now delve deeper to investigate questions about transcriptome complexity, sample complexity, and a variety of questions in RNA biology beyond transcription.

References

1. Alwine JC, Kemp DJ, Stark GR. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc Natl Acad Sci U S A. 1977;74(12):5350–4.
2. Wang AM. Quantitation of mRNA by the polymerase chain reaction. Proc Natl Acad Sci U S A. 1989;86(24):9717–21.
3. Adams M, Kelley J, Gocayne J, Dubnick M, Polymeropoulos M, Xiao H, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. Science. 1991;252(5013):1651–6.
4. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270(5235):467–70.
5. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008;24(3):133–41.

6. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20 (11):631–56.
7. Illumina Sequencing Platforms [Internet]. Illumina, Inc. [cited 2021 Jan 5]. https://www.illumina.com/systems/sequencing-platforms.html.
8. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol. 2014;32(9):903–14.
9. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12(10): 671–82.
10. Borodina T, Adjaye J, Sultan M. A strand-specific library preparation protocol for RNA sequencing. Methods Enzymol. 2011;500:79–98. https://linkinghub.elsevier.com/retrieve/pii/B9780123851185000050.
11. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17(1):13.
12. Sambrook J, Russell DW. Purification of RNA from cells and tissues by acid phenol-guanidium thiocyanate-chloroform extraction. CSH Protoc. 2006;2006(1):pdb.prot4045.
13. Vanecko S, Laskowski M Sr. Studies of the specificity of deoxyribonuclease I: III. Hydrolysis of chains carrying a monoesterified phosphate on carbon 5′. J Biol Chem. 1961;236(12):3312–6.
14. Sheng Q, Vickers K, Zhao S, Wang J, Samuels DC, Koues O, et al. Multi-perspective quality control of Illumina RNA sequencing data analysis. Brief Funct Genomics. 2017;16(4):194–204. https://doi.org/10.1093/bfgp/elw035.
15. Gallego Romero I, Pai AA, Tung J, Gilad Y. RNA-seq: impact of RNA degradation on transcript quantification. BMC Biol. 2014;12(1):42.
16. Evers DL, He J, Kim YH, Mason JT, O'Leary TJ. Paraffin embedding contributes to RNA aggregation, reduced RNA yield, and low RNA quality. J Mol Diagn. 2011;13(6):687–94.
17. Barbas CF III, Burton DR, Scott JK, Silverman GJ. Quantitation of DNA and RNA. CSH Protoc. 2007;2007:pdb.ip47.
18. Eddy SR. Non-coding RNA genes and the modern RNA world. Nat Rev Genet. 2001;2(12): 919–29.
19. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol. 2006;7 (1):3.
20. Dreyfus M, Régnier P. The poly(A) tail of mRNAs. Cell. 2002;111(5):611–3.
21. Kung JTY, Colognori D, Lee JT. Long noncoding RNAs: past, present, and future. Genetics. 2013;193(3):651–69.
22. Cabanski CR, Magrini V, Griffith M, Griffith OL, McGrath S, Zhang J, et al. cDNA hybrid capture improves transcriptome analysis on low-input and archived samples. J Mol Diagn. 2014;16(4):440–51.
23. Curion F, Handel AE, Attar M, Gallone G, Bowden R, Cader MZ, et al. Targeted RNA sequencing enhances gene expression profiling of ultra-low input samples. RNA Biol. 2020;17 (12):1741–53.
24. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: overviews and challenges. Biotechniques. 2014;56(2):61–4. https://www.future-science.com/doi/10.2144/000114133.
25. Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis: RNA-Seq. WIREs RNA. 2017;8(1):e1364.
26. Pelechano V, Steinmetz LM. Gene regulation by antisense transcription. Nat Rev Genet. 2013;14 (12):880–93.

27. Dominic Mills J, Kawahara Y, Janitz M. Strand-specific RNA-Seq provides greater resolution of transcriptome profiling. Curr Genomics. 2013;14(3):173–81.
28. Illumina, Inc. An introduction to next-generation sequencing technology. San Diego: Illumina; 2017.
29. Chen F, Dong M, Ge M, Zhu L, Ren L, Liu G, et al. The history and advances of reversible terminators used in new generations of sequencing technology. Genomics Proteomics Bioinform. 2013;11(1):34–40.
30. Illumina, Inc. 2-channel SBS Technology. San Diego: Illumina; 2021. https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology/2-channel-sbs.html.
31. Illumina, Inc. Illumina CMOS chip and one-channel SBS chemistry, vol. 4. San Diego: Illumina; 2018.
32. Illumina, Inc. Paired-end vs. single-read sequencing technology. San Diego: Illumina; 2021. https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html.
33. Kukurba KR, Montgomery SB. RNA sequencing and analysis. Cold Spring Harb Protoc. 2015;2015(11):pdb.top084970.
34. Moreton J, Izquierdo A, Emes RD. Assembly, assessment, and availability of de novo generated eukaryotic transcriptomes. Front Genet. 2016;6:361. http://journal.frontiersin.org/Article/10.3389/fgene.2015.00361/abstract.
35. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016;11(9):1650–67.
36. Morlan JD, Qu K, Sinicropi DV. Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. PLoS One. 2012;7(8):e42882.
37. Atkinson SR, Marguerat S, Bähler J. Exploring long non-coding RNAs through sequencing. Semin Cell Dev Biol. 2012;23(2):200–5.
38. Li J, Liu C. Coding or noncoding, the converging concepts of RNAs. Front Genet. 2019;22(10):496.
39. Choudhuri S. Small noncoding RNAs: biogenesis, function, and emerging significance in toxicology. J Biochem Mol Toxicol. 2010;24(3):195–216.
40. Liu Q, Ding C, Lang X, Guo G, Chen J, Su X. Small noncoding RNA discovery and profiling with sRNAtools based on high-throughput sequencing. Brief Bioinform. 2019;22(1):463–73.
41. Zhang P, Wu W, Chen Q, Chen M. Non-coding RNAs and their integrated networks. J Integr Bioinform. 2019;16(3):20190027. https://www.degruyter.com/view/journals/jib/16/3/article-20190027.xml.
42. Dard-Dascot C, Naquin D, d'Aubenton-Carafa Y, Alix K, Thermes C, van Dijk E. Systematic comparison of small RNA library preparation protocols for next-generation sequencing. BMC Genomics. 2018;19(1):118.
43. Calviello L, Ohler U. Beyond read-counts: Ribo-seq data analysis to understand the functions of the transcriptome. Trends Genet. 2017;33(10):728–44.
44. Wissink EM, Vihervaara A, Tippens ND, Lis JT. Nascent RNA analyses: tracking transcription and its regulation. Nat Rev Genet. 2019;20(12):705–23.
45. Ganser LR, Kelly ML, Herschlag D, Al-Hashimi HM. The roles of structural dynamics in the cellular functions of RNAs. Nat Rev Mol Cell Biol. 2019;20(8):474–89.
46. Strobel EJ, Yu AM, Lucks JB. High-throughput determination of RNA structures. Nat Rev Genet. 2018;19(10):615–34.
47. Wang B, Kumar V, Olson A, Ware D. Reviving the transcriptome studies: an insight into the emergence of single-molecule transcriptome sequencing. Front Genet. 2019;26(10):384.
48. Rhoads A, Au KF. PacBio sequencing and its applications. Genomics Proteomics Bioinform. 2015;13(5):278–89.

49. Kono N, Arakawa K. Nanopore sequencing: review of potential applications in functional genomics. Develop Growth Differ. 2019;61(5):316–26.
50. Piovesan A, Antonaros F, Vitale L, Strippoli P, Pelleri MC, Caracausi M. Human protein-coding genes and gene feature statistics in 2019. BMC Res Notes. 2019;12(1):315.
51. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al. The genome sequence of Drosophila melanogaster. Science. 2000;287(5461):2185–95.
52. Miura F, Kawaguchi N, Yoshida M, Uematsu C, Kito K, Sakaki Y, et al. Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs. BMC Genomics. 2008;29(9):574.
53. Schwartz S, Motorin Y. Next-generation sequencing technologies for detection of modified nucleotides in RNAs. RNA Biol. 2017;14(9):1124–37.
54. Xu L, Seki M. Recent advances in the detection of base modifications using the nanopore sequencer. J Hum Genet. 2020;65(1):25–33.
55. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018;50(8):96.