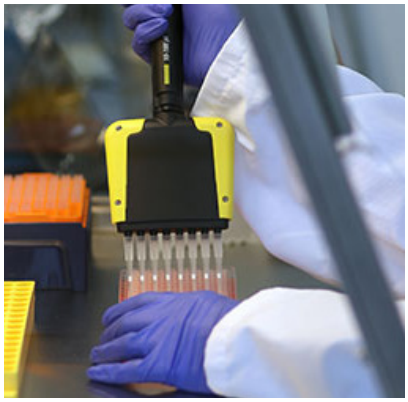
LEXOGEN
The RNA Experts

RNA LEXICON

CHAPTER #7

RNA-Seq Library Preparation: Molecular Biology Basics

After you have successfully extracted the RNA from your sample, controlled the quality of your preparation, and removed residual gDNA (if needed), it is time to prepare your RNA-Seq libraries. Depending on the library method you have chosen and the RNA fraction you are interested in, you may need to pre-select for your RNA fraction of choice. For more details on this, check out our chapter on RNA Enrichment and Depletion.



While the individual reaction steps in an RNA-Seq workflow can vary depending on the library preparation method used, the molecular principles underlying these reaction steps remain the same. The following reactions are commonly used in RNA-Seq library preparation: Reverse Transcription, Second Strand Synthesis, End Repair, Ligation, and PCR Amplification. In the following chapter, we will go over these steps and shed light on the molecular basis of these reactions.

1. Reverse Transcription / First Strand Synthesis

First strand synthesis refers to the generation of a complementary DNA molecule from an RNA template by an enzyme called Reverse Transcriptase. Reverse transcriptases are a viral RNA-dependent DNA-polymerases that transcribe RNA template molecules into copy DNA (cDNA). The discovery of reverse transcriptases in the 1970s [1,2] has revolutionized molecular biology by short circuiting Francis Cricks' original dogma of molecular biology which described a unilateral flow of genetic information from DNA → RNA → Protein (Figure 1). This discovery was awarded with a Nobel Prize in 1975.

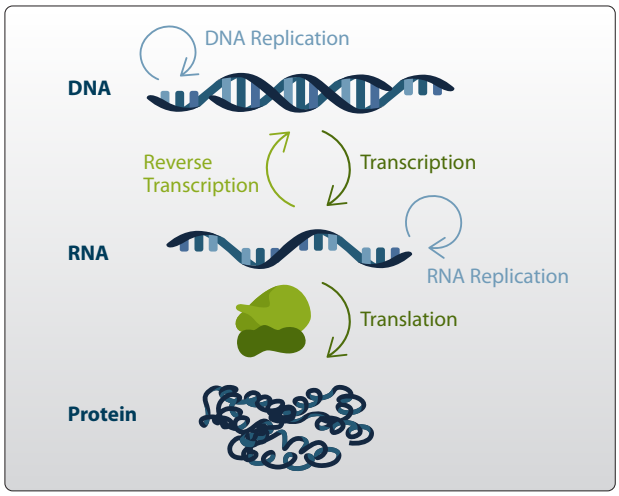


Figure 1 | Central Dogma of Molecular Biology.

The central dogma of molecular biology describes the flow of genetic information within biological systems. DNA can either be replicated by DNA Polymerase (DNA →DNA) or is transcribed into RNA by RNA Polymerases (DNA → RNA). RNA is a messenger molecule that can be translated into proteins by Ribosomes (RNA → protein). Reverse transcription describes the conversion of RNA into DNA (RNA → DNA) by reverse transcriptase and is commonly used by viruses. RNA replication generates RNA from RNA.

Today, reverse transcription is at the core of all RNA-Seq experiments and as such remains one of the fundamental reactions in state-of-the-art molecular biology applications.

A reverse transcription reaction in an RNA-Seq workflow requires three sub-steps: primer annealing, cDNA synthesis (Fig. 2), and enzyme inactivation.

Primer Annealing

An oligonucleotide primer is hybridized to a complementary sequence within the RNA template molecule. This step usually begins with a short incubation at a higher temperature that opens up structures in the RNA and is followed by a cool down to a lower temperature during which the primers hybridize to the RNA. Depending on the nature of the primer, the appropriate annealing strategy should be chosen.

For example, commonly used short random primers (e.g., hexamers) have a rather low annealing temperature requirement. For longer target-specific primers that bind to a defined sequence in a transcript of interest, a much higher annealing temperature is used. This prevents unspecific priming to sequences with partial complementarity. Similarly, oligo(dT) primers which hybridize to poly(A) sequences and are thus commonly used in mRNA-Seq and 3'-Seq workflows should ideally be kept at the reaction temperature to avoid mis-priming, e.g., to shorter A-rich sequences that can be located within rRNA transcripts.

cDNA Synthesis

In this step, the oligonucleotide primer is elongated by the reverse transcriptase (Fig. 2). The enzyme binds and adds the complementary nucleotide (dNTP) based on the sequence of the RNA template strand. Depending on the primer used for reverse transcription, the reaction is either directly kept at an incubation temperature of 37 – 50 °C or may require a short incubation at a lower temperature.

Due to their low melting temperature, short random primers dissociate from the target molecule at elevated temperatures. Therefore, a short period for primer extension at a lower temperature (e.g., 25 °C) is used to increase the length of the hybridized sequence. The elongated primer then stays attached to the RNA template when the reaction temperature is elevated after the pre-incubation step. This procedure increases the efficiency of the complete reaction.

The reaction temperature is further dependent on the enzyme that is used. As reverse transcriptases are derived from

viruses and are evolved for cDNA generation inside of host organisms, most of these enzymes have an optimum temperature of 37 – 42 °C. Thermo stabile enzymes have been engineered for cDNA synthesis from highly structured and demanding transcripts.

As reverse transcriptases lack proofreading function (3′ → 5′ exodeoxyribonuclease activity), the rate of nucleotide misincorporation is higher than for DNA-Polymerases and estimated at around 1 in 1,000 to 10,000 bases.

For short-read sequencing workflows, these mutations are rare as the fragments generated are commonly between 50 – 500 nucleotides and the analysis of replicates can help identify mutations introduced by reverse transcription.

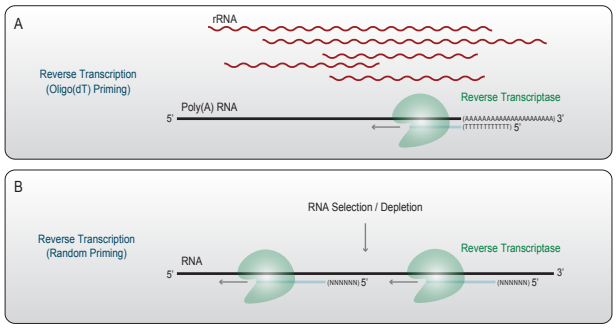


Figure 2 | Reverse Transcription generates cDNA from RNA template molecules. A) Oligo(dT)-primed reverse transcription does not require rRNA depletion or mRNA selection due to the primer annealing specifically to the 3′ poly(A) tail of mRNAs. B) Random-primed reverse transcription uses pre-selected or depleted input RNA. Random primers hybridize along the RNA template. Reverse transcriptase elongates the primers and generates a complementary DNA copy.

Enzyme Deactivation / Removal

The last part of the reaction is the deactivation or removal of the reverse transcriptase to avoid interference with downstream reaction steps. This can either be done by a short heating step at 70 – 85 °C or using a clean-up step.

2. RNA Template Removal and Second Strand Synthesis

After the first DNA strand is generated, many RNA-Seq workflows require the generation of a double stranded DNA molecule. The product of the reverse transcription reaction is a cDNA single strand that is paired with the initial RNA template strand. In order to generate the second DNA strand, the RNA first needs to be removed. There are different ways to achieve RNA removal. Most commonly, the product is heated in a buffer that is formulated to specifically hydrolyze RNA while the DNA strand stays intact.

Subsequently, random primers are annealed to the now accessible cDNA first strand and the second strand is generated by incorporation of complementary nucleotides using a DNA-dependent DNA Polymerase. For random primed second strand synthesis, DNA-Polymerases that work at lower temperatures are used, while their thermo stable counterparts are mostly used in PCR amplification and for targeted RNA-Seq approaches.

RNA template removal is required to ensure that short random primers used to initiate second strand synthesis gain access to the cDNA first strand. In case a targeted sequencing approach is used, RNA removal can be omitted. Targeted primers are commonly designed with a high annealing temperature (>60 °C) to avoid unspecific priming to non-target sequences. Concomitantly, the reaction is carried out at a much higher temperature than random primed second strand synthesis. This elevated temperature of 60 – 72 °C weakens the RNA-DNA interactions sufficiently for the targeted primer to anneal to the cDNA first strand. The RNA template is unwound and the complementary DNA second strand is generated.

Most DNA-polymerases used for second strand synthesis possess proofreading activity and thus the error rates in this step are much lower than during reverse transcription.

The properties of natural reverse transcriptases in combination with a process called nick translation for second strand synthesis were already exploited in the 1980s for efficient generation of cDNA libraries [3].

Second Strand Synthesis by Nick Translation

Another method for second strand generation uses the properties of reverse transcriptases with RNase H activity. This activity is present in many wild-type versions of reverse transcriptase but has been deactivated in enzymes routinely used for RNA-Seq approaches to preserve the template RNA. RNase H activity cleaves the RNA template molecule once it is paired with the complementary cDNA strand providing so called nicks. The remaining short RNA pieces stay attached to the cDNA first strand and can be extended by DNA-Polymerase synthesizing short complementary second strands. During this process, the short RNA strands are replaced by the repair function (5′ → 3′ exonuclease activity) of the DNA-Polymerase, in a process called nick translation. The remaining breaks between the individual pieces are then sealed by a DNA Ligase.

Second strand synthesis by nick translation utilizes the inherent repair function of DNA-Polymerase I from *Escherichia coli*. This enzyme is active in the replication of the *E. coli* chromosome and possesses a repair function, i.e., 5′ → 3′ exonuclease activity which is different from the proofreading activity. In the bacterium, this function is used to repair single strand breaks in the genomic DNA that compromise genome integrity and impair transcription.cated within rRNA transcripts.

3. End Repair

First and Second Strand Synthesis generate partially double stranded DNA with single stranded ends. A process termed end repair is therefore often used to prepare the double-stranded DNA library for the adapter ligation step. During the end repair reaction, partial single strands on the 5′ end of the fragment are filled in by a polymerase to generate a double strand using the protruding end as a template. Single stranded 3′ overhangs on the other end of the DNA fragment are removed using a 3′ → 5′ exonuclease. This process creates blunt ends on both sides of the fragment.

Depending on the adapter ligation strategy, a single adenine can be added at the 3′ end of each strand in a process referred to as A-tailing. These A-tailed fragments are subsequently ligated to adapters with a single 5′ T-overhang in the subsequent ligation step (Fig. 3).

As the efficiency of blunt end ligation is usually lower than ligation with overhangs (even if it is just one nucleotide), end repair including A-tailing is a common theme in library preparation for Next Generation Sequencing.

As DNA-Ligases require molecules with a 5′ phosphate (5′-P) and a 3′ hydroxyl group (3′-OH) as substrates, these functional groups are also generated during the end repair process. This is achieved either by using enzymes that generate such end products or by enzymatic activities that transfer these functional groups, e.g., Polynucleotidekinase (PNK) can transfer phosphate groups to the 5′ end of RNA.

There are also other variations of "end repair" used in RNA-Seq library preparation. The common theme is that overhangs are removed or filled in. This is mostly done on DNA, but RNA can also undergo end repair.

4. Ligation

We already outlined a few basic principles of ligation in the previous section, here we will go into a few more details. Ligation in molecular biology refers to the enzymatic process of joining two nucleic acid molecules by attaching the 3′-OH group of the first molecule to the 5′-P group of the second molecule. Ligases were discovered by multiple labs [4] in the 1960s and are also considered as one of the major breakthroughs in molecular biology, enabling molecular cloning of recombinant DNA molecules and NGS library generation.

As mentioned above, ligation occurs in various modes.

Blunt-end Ligation

Blunt-end ligation refers to the joining of two double stranded DNA molecules without any overhangs. The ligation depends on the random collision of the two molecules to be joined and is thus much less efficient than ligation of molecules with complementary overhangs. When 5′ phosphates and 3′ hydroxyl groups are present in the same molecule, self-ligation leading to circularization of just one reaction partner is favored, as the ends of the same molecule are per default in close proximity and thus more likely to collide. Blunt-end ligation is generally avoided in NGS library preparation due to the lower efficiency.

Sticky-end Ligation / TA Ligation

Sticky-end ligation occurs on DNA fragments with compatible single stranded overhangs on the two molecules to be ligated. These complementary overhangs, also called cohesive ends, can anneal between the two molecules and thus increase the efficiency of ligation. TA ligation is a special form of sticky-end ligation with an overhang of only one nucleotide. As the efficiency is increased dramatically for TA ligation, this process is commonly used in NGS library preparation. A single adenine is added to the 3′ end of each strand of the double-stranded DNA insert during A-tailing and the inserts are then ligated with partially double stranded sequencing adapters carrying a protruding T at the corresponding 5′ ends (Fig. 3).

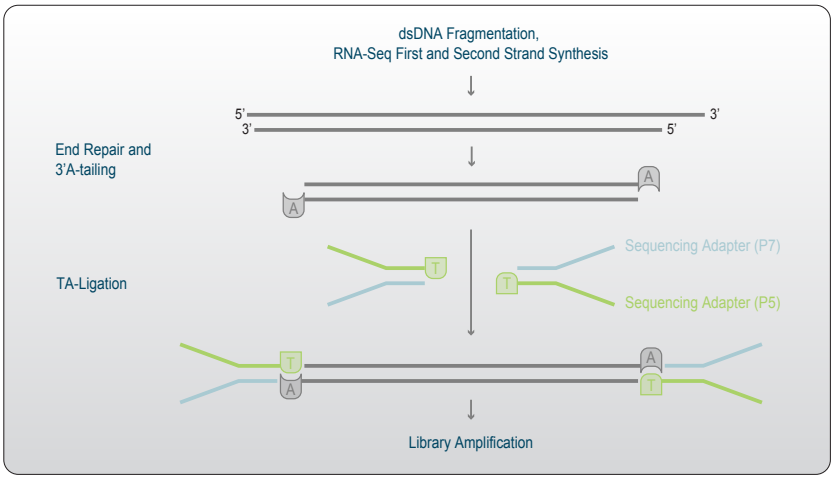


Figure 3 | End Repair, A-tailing and TA ligation. After first and second strand synthesis, single stranded overhangs are removed and the 3′ ends are adenylated. Pre-annealed partially double-stranded sequencing adapters with 3′ T-overhangs are ligated to the A-tailed inserts. Sequencing adapters consist of a P7 and a P5 linker sequence.

Single-stranded ligation

Some NGS library preparation methods also use single-stranded ligation. Here, two single-stranded nucleic acids are covalently linked. These molecules can be DNA, RNA, or a combination of both. Single-stranded ligation can be performed directly on cDNA first strands to introduce partial sequencing adapter and thereby omit the need for a dedicated second strand synthesis. The more common use case for single-stranded ligation, however, is the ligation of adapters to short RNA molecules, especially during library generation for small RNA sequencing.

Small RNAs can be as short as 15 nucleotides and therefore cannot be picked up by random-primed library prep methods. For generation of small RNA libraries, sequencing adapters are ligated to the 3' and the 5' end of the short RNA molecules. Primers with complementarity to the 3' adapter sequence are then used to initiate reverse transcription and convert the ligation product to an NGS library.

5. PCR Amplification

PCR or polymerase chain reaction is a widely used method to generate millions of DNA copies from as low as a single molecule. The description of the PCR reaction [5] and its first application in the 1980s is mainly attributed to Kary Mullis who received the 1993 Nobel Prize in Chemistry for his discoveries. The extreme sensitivity and versatility of PCR led to its numerous applications in various fields of science ranging from molecular biology research to medical diagnostics, to infectious disease detection (e.g., SARS-CoV-2 diagnostics), even down to paternity testing and criminal forensics, and has helped to unravel the genome of Neanderthals.

Thus, polymerase chain reaction is an integral part of cutting-edge science and used in many state-of-the-art techniques, including high-throughput Next Generation Sequencing. PCR is the final step in most NGS library preparation workflows. During this step, the libraries are amplified for quality control. In case partial adapters were introduced in the library generation step, the adapter sequences are completed and indices are introduced.

PCR uses a thermostable polymerase to amplify a DNA template by repeating three steps in multiple cycles: denaturation, primer annealing, and elongation.

PCR cyclers for running three different PCR reactions in one machine.

Denaturation

The DNA template is heated to 94 – 99 °C to melt the DNA double helix and separate the double strands into single strands. Denaturation also serves to release the DNA-polymerase from the DNA molecule that was completed in the previous cycle.

The denaturation temperature depends on the polymerase that is used during the reaction. The first thermostable polymerase that was identified and used in PCR is *Taq*-Polymerase, derived from *Thermus aquaticus*, a bacterium found in hot springs and hydrothermal vents. *Taq*-Polymerase is still widely used in PCR reactions and can withstand several short denaturation rounds at 94 °C. As Taq-Polymerase lacks proofreading activity, demanding applications including sequencing commonly use specifically engineered high-fidelity polymerases instead of *Taq*. These polymerases possess proofreading activity and synthesize new DNA molecules extremely fast due to mutations introduced into their DNA binding domain. These mutations increase the processivity of the enzyme by strengthening their ability to bind the template DNA for a longer duration. As a result of more efficient binding, a higher denaturation temperature of up to 99 °C is needed to remove the polymerase efficiently and start the new cycle. If the denaturation temperature is too low, the enzyme is not efficiently released after each cycle and the amplification is impaired.

Primer Annealing

During this step, the reaction temperature is decreased, and primers anneal to complementary sequences of the single stranded DNA template. The forward primer is thereby annealed to the sense strand, the reverse primer binds to a complementary sequence of the antisense strand. The sequence that is amplified during PCR is the sequence encompassed by the primer pair. Thus, for amplification of NGS libraries, one of the primers is complementary to the (partial) P5 adapter and the second primer is complementary to the (partial) P7 adapter.

Primer annealing is crucial for a successful PCR; therefore, it is important to determine the proper annealing temperature. The annealing temperature depends on the primer sequence and needs to allow the primers to hybridize to the template specifically. If the annealing temperature is set too high (above the melting temperature of one or both primers), the primer may not bind at all. It is similarly detrimental to use an annealing temperature that is too low, as the primers can also bind imperfectly to sequences with only partial complementarity, generating undesired by-products. Typically, annealing temperatures are between 3 – 5 °C lower than the melting temperature (T_m) of the primers. As the primer with the lower T_m ultimately determines the annealing temperature of the complete reaction, primers should be designed in a way that the melting and annealing temperatures are closely matched for the pair that should be used.

Elongation

During this step, the DNA polymerase catalyzes the incorporation of nucleotides complementary to the DNA template strand, i.e., the primers are elongated. The complementary strand is generated by polymerization, i.e., by enzymatically linking the 5' phosphate of the deoxyribonucleotide triphosphates (dNTPs) to the OH group at the 3' end of the newly synthesized strand. During elongation or extension phase temperatures between 65 – 72 °C are used, corresponding to optimal temperature of the enzyme used. Elongation time depends on the length of the fragment that shall be amplified and the polymerase used.

For example, as a rule of thumb, *Taq*-Polymerase requires approximately 90 seconds to synthesize a DNA fragment of 1 kb length. In contrast, the highly processive engineered polymerases mentioned above can synthesize fragments of up to 3 kb in just 30 seconds. For PCR amplification of NGS libraries, elongation times of ~30 seconds to 1 minute are used.

After the complementary strands are synthesized during elongation, the process is repeated again starting with denaturation. With each new cycle, the original DNA template and all copies generated in the previous cycles are available for primer annealing and elongation. Thus, the template is amplified exponentially, i.e., the number of molecules essentially doubles with each cycle. When the desired level of amplification is reached, and the reaction is stopped, and a final elongation converts all partial single strands into double-stranded DNA.

Upon consumption of reaction components, such as dNTPs and primers, and gradual loss of polymerase activity, the reaction slows and eventually enters a plateau where products no longer accumulate.

Once the plateau is reached and primers are consumed, complementary regions in the templates themselves can base pair and generate partially double-stranded regions. Due to uneven distribution of bases in the amplified fragment, some nucleotides may be depleted to a larger extent. Thus, errors can accumulate in the later cycles due to nucleotide depletion and loss of polymerase fidelity.

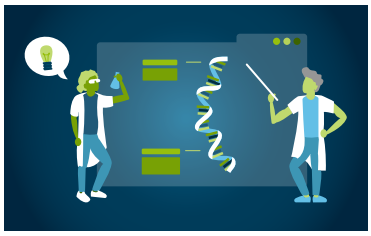
One important aspect for PCR amplification, especially during NGS library preparation, is to avoid entering the plateau stage. Overcycling, i.e., the application of too many PCR cycles, can have a negative impact on NGS data quality, as outlined in Chapter 2 focusing on the sequencing process.

Literature:

1. Temin HM, Mizutani S (1970) RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* 226(5252):1211–1213, DOI: 10.1038/2261211a0
2. Baltimore D (1970) RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* 226(5252):1209–1211, DOI: 10.1038/2261209a0
3. Gubler U, Hoffman BJ (1983) A simple and very efficient method for generating cDNA libraries. *Gene* 25(2-3):263-269, DOI: 10.1016/0378-1119(83)90230-5
4. Lehman IR. DNA ligase: structure, mechanism, and function. Science. 1974 Nov 29;186(4166):790-7, PMID: 4377758, DOI: 10.1126/science.186.4166.790
5. Mullis, K.F.; Faloona, F.; Scharf, S.; Saiki, R.; Horn, G.; Erlich, H. (1986). "Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction". *Cold Spring Harbor Symposia on Quantitative Biology*. 51: 263–273. DOI:10.1101/sqb.1986.051.01.032

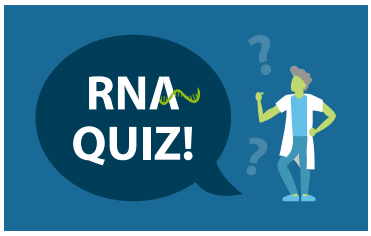
LEXOGEN
The RNA Experts

Curious to learn more?



Explore more chapters in our RNA LEXICON:
www.lexogen.com/rna-lexicon

Watch our RNA EXPERTise Videos:
www.lexogen.com/rna-expertise-videos



Show your RNA expertise and master all questions of our RNA Quiz:
www.lexogen.com/lexicon-quiz-3

Lexogen GmbH
Campus Vienna Biocenter 5
1030 Vienna, Austria
Telephone: +43 (0) 1 345 1212
Fax: +43 (0) 1 345 1212-99
info@lexogen.com
www.lexogen.com

Lexogen, Inc.
51 Autumn Pond Park
Greenland, NH 03840, USA
Telephone: +1-603-431-4300
Fax: +1-603-431-4333