LEXOGEN
The RNA Experts

RNA LEXICON

eBOOK

Welcome
to the world of RNA

Table of Contents

Introduction to Next-Generation RNA Sequencing

Next-Generation Sequencing (NGS) is the gold-standard for genomic and transcriptomic research in life science fields . While it shares certain core similarities with methods such as Sanger sequencing, the absolutely massive level of throughput made possible by NGS sets it worlds apart and has revolutionized the way scientists work.

RNA Sequencing (RNA-Seq) specifically is at the cutting edge of NGS capabilities. In its simplest form, RNA-Seq allows us to determine RNA molecules in a sample at the moment of sampling. The transcriptome is a highly dynamic cellular feature that opens up a world of discovery potential. Changes in response to drugs, various states of disease, post-transcriptional modifications, and alternatively spliced transcripts are just some examples of discoveries made possible by RNA-Seq.

In short, RNA-Seq allows us to assess the whole transcriptome at unprecedented levels of sensitivity and generate a snapshot of the transcriptome under conditions and points in time specific to the topic being studied. Though this serves as a perfectly suitable and brief summary of the broader field of RNA-Seq, there are many "flavors" of RNA-Seq to try.

Short-read sequencing is performed commonly on RNA-seq library preps. Sequencing is typically done on shorter inserts generated from the sample material between 50 – 500 bp in length and then re-assembled or counted in downstream data analysis. Long-read sequencing is frequently used as a DNA-Seq method as well as an RNA-Seq method and can determine the sequence of the sample at hand at a high capacity, e.g., between 10,000 – 100,000 base pairs at a time. For RNA-Sequencing, combining long-read and short-read sequencing is especially useful to determine the transcriptomes of unknown or under annotated species. Long-read sequencing can determine the exact transcript sequence and provide a scaffold on which the short reads can be assembled similar to puzzle pieces.

Bulk RNA-Seq measures the gene expression of a set of samples without differentiating between the cell types within the sample. This method provides a broad overview of gene expression in a set of samples. The range of RNA-Seq applications is extremely broad, and the boundaries are being pushed with every passing day (see the List below for example applications). With such a wide range of applications, it is no surprise that RNA-Seq is a useful, powerful tool that offers quantitative, qualitative, and time-resolved data on set of experimental samples.

List of Common RNA-Seq Applications

- ✓ alternative polyadenylation events,
- ✓ alternative splicing analysis,
- ✓ biomarker discovery,
- ✓ cell type and subtype characterization,
- ✓ differential expression,
- ✓ discovery of novel genes and gene isoforms,
- ✓ expression quantification,
- ✓ host / pathogen interaction (dual RNA-seq),
- ✓ identification of genuine transcription start sites (and promoters),
- ✓ metatranscriptomics (community transcriptomes),
- ✓ precise 3' end mapping,
- ✓ RNA kinetics (biosynthesis and decay),
- ✓ tissue-specific gene expression,
- ✓ transcriptome assembly of uncharacterized species, and many other applications.

Here at Lexogen we are focused on the short-read RNA sequencing workflow. While there are exceptions to this, the general outline of an experimental procedure is as follows:

1. RNA is isolated from the sample and contaminating DNA is removed, e.g., with DNase treatment.
2. If needed for the method, the RNA is pre-treated, i.e., the mRNA is enriched by polyA selection or rRNA depletion.

   ! Very commonly a 3' mRNA library generation method is used which enriches for poly(A) tail-containing material, as this region is rich in features and enables sound differential gene expression studies without over-working the samples ahead of time.
3. RNA can then be fragmented. If using fragmentation-free protocol, this point can be ignored.

4. Reverse transcription is performed using the RNA sample and library generation primers as chosen.
5. A second strand synthesis is then performed by randomly priming along the first cDNA strand.
6. At this point the libraries are ready for end repair and the ligation of any products occurs on either double stranded cDNA.
7. PCR is finally performed to add indices to the library, and / or amplify the material for library quality control.

Lexogen library preps for mRNA-Seq and whole transcriptome sequencing use a slightly different approach. Partial adapters are introduced already in the first step, e.g., during reverse transcription. Thereby, workflows are streamlined and efficiently shortened by removing various processing steps. This saves valuable time and allows to complete the library preparation workflow within a few hours. Figure 1 illustrates the schematic workflow for two different RNA-Seq library preparation methods.

RNA-Seq can be used for quite a variety of applications, is highly customizable by the combination of different methods, and offers sheer endless possibilities for modifications to fit the individual needs of a research project. Planning your experiment before taking up the pipets is the key to success and ensures high quality results to foster your ideas and hypotheses.

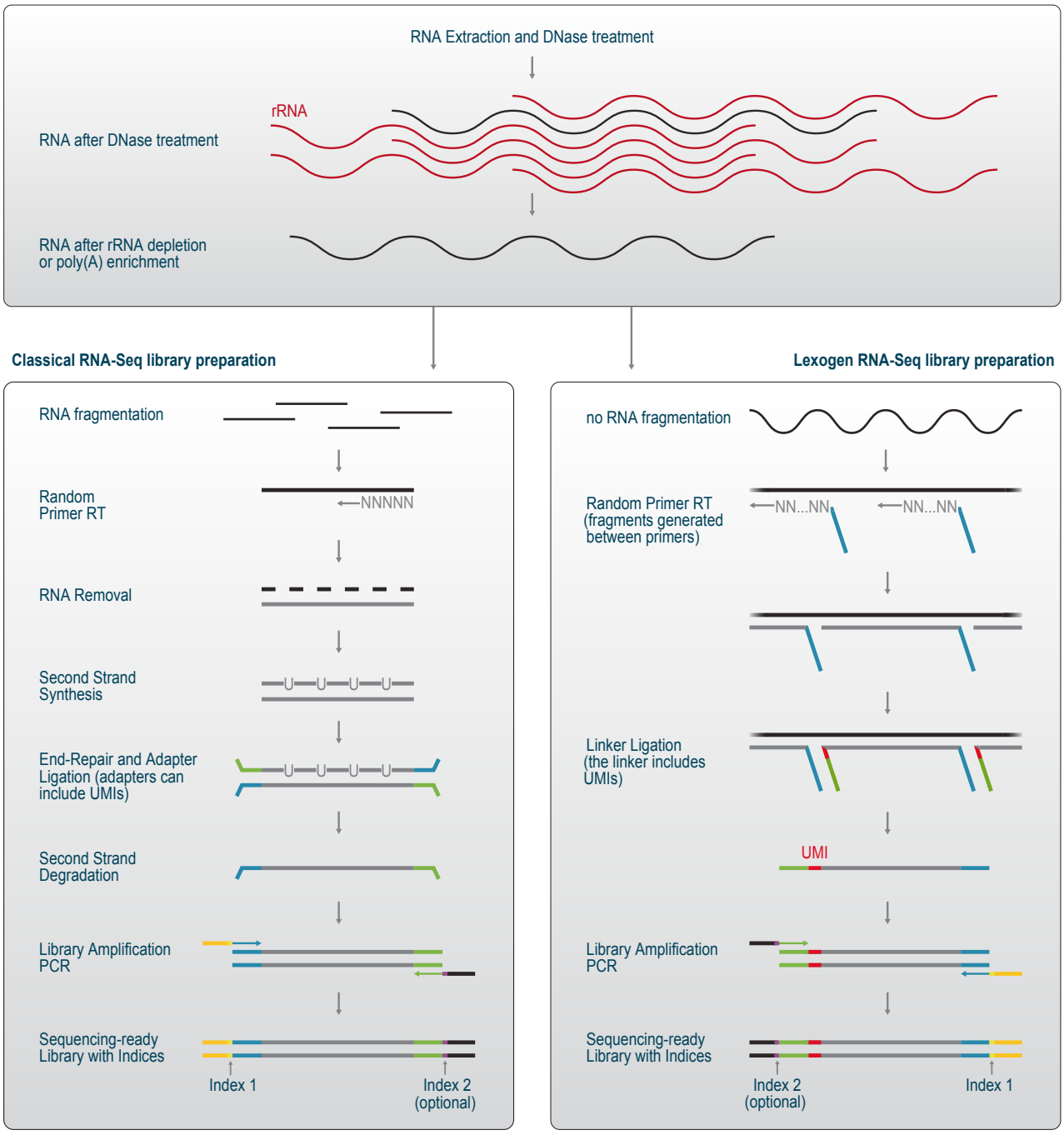


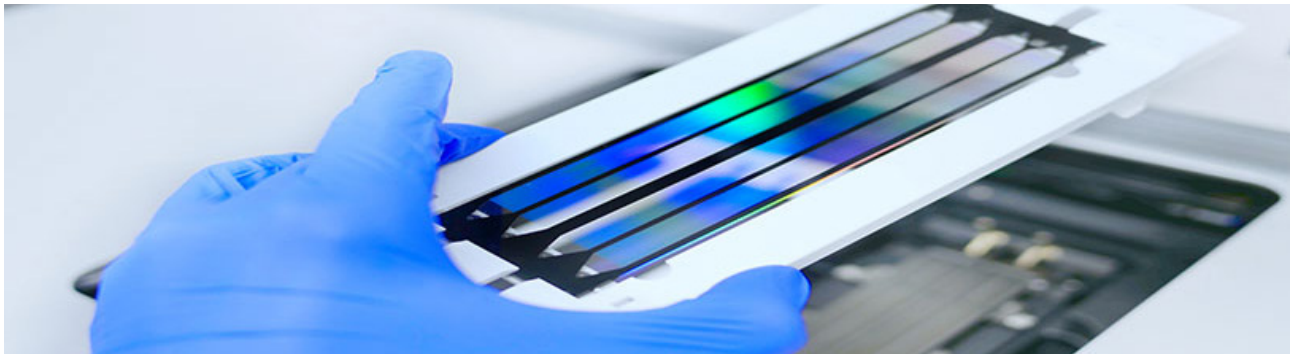
Figure 1 | RNA-Seq Library Preparation Workflows for Illumina short-read sequencing. RNA is extracted from cells, tissues, bioliquids or other samples and treated with DNase to remove genomic DNA prior to depletion of ribosomal RNAs or poly(A) RNA enrichment.

Left: Schematic overview depicting classical RNA-Seq workflows. Pre-treated RNA is fragmented, and cDNA is generated by random primed reverse transcription. Following RNA removal, the second strand is generated. Labeling with dUTP allows to generate libraries that retain strandedness. Double-stranded cDNA is end-repaired, sequencing adapters are ligated, and the dUTP-labeled strand is specifically degraded. Thereby, only one strand remains, and the strand information is retained. PCR is performed on this remaining strand to amplify the library, complete the adapter sequences for sequencing, and introduce indices.

Right: Lexogen RNA-Seq library preparation protocols omit RNA fragmentation and cDNA first strands are generated from random primers that already contain partial adapter sequences. Inserts are generated between two hybridized primers, and the second partial adapter is ligated to the cDNA first strands, thus retaining strandedness. The linker also contains Unique Molecular Identifiers (UMIs) which will be discussed in a separate chapter. Sequencing ready libraries containing UMIs, full-length adapters and indices are then obtained by PCR.

Next Generation Sequencing: How "Sequencing by Synthesis" Works

At Lexogen we are designing and producing RNA-Sequencing library preparation kits for use on Illumina sequencing instruments. How RNA-Seq libraries can be generated is described in Chapter 1, our Introduction to RNA Sequencing. In the following chapter we will focus on the sequencing process itself and its underlying principles. In order to interpret library quality parameters, it is helpful to know how the sequencing process known as "Sequencing by Synthesis" functions.



1. Preparing the Library

After the libraries are generated, amplified by PCR, and passed quality control, they are prepared for sequencing. During this process, the concentration is adjusted to the requirements of the sequencer and the double-stranded libraries are denatured to single strands as only single strands can be bound onto the flow cell. Ready-to-sequence libraries contain specific Illumina adapter sequences, termed P5 and P7, at their 5' and 3' end (Fig. 1).

These adapter sequences serve two functions:

1. The "outer" region (shown in black and orange) is required for binding to complementary sequences on the surface of the Illumina flow cell. Thereby, the individual library single strands are captured to be sequenced.
2. The "inner" region (shown in green and blue) serves as binding site for the sequencing primer which is used to read out the insert sequence during the actual sequencing process.

*Figure 1 | Structure of a sequencing-ready Illumina-compatible library. The insert sequence (gray) is flanked by two sequencing adapters. The P5 adapter contains a flow cell binding region (black). This sequence can also coincide with the binding site for the Index 2 sequencing primer for the optional i5 index (Index 2, purple). * Depending on the sequencer the index 2 sequencing primer binding site can be located in the inner or outer region of the adapter. The P5 adapter also contains the Read 1 sequencing primer binding site (green). The P7 adapter contains a flow cell binding region (orange), the i7 index sequence (Index 1, yellow) and the Read 2 / Index 1 sequencing primer binding sites (blue).*

2. Cluster Generation

In the first step, the single-stranded sequencing-ready libraries are loaded onto the flow cell. The complementary oligonucleotides on the flow cell surface act as an anchor to capture the libraries by binding to the outer region of the Illumina adapter sequence. Once attached, clusters can be generated, and libraries are sequenced by synthesis.

Cluster generation begins by bridge amplification. During this process, the complementary strand is generated by elongating the oligo attached to the flow cell. Thereby, a copy of the molecule to be sequenced is now covalently attached to the flow cell. The original molecule is washed away, and the strand bends over (like a "bridge") to attach to the next flow cell oligo. This second oligo is complementary to the other sequencing adapter, and thus upon elongation the reverse strand is generated. After forward and reverse strands are generated and both are stably attached to the flow cell, clusters are generated by clonal amplification, i.e., the process is repeated over and over until the required level of amplification is reached. Essentially, enough material needs to be generated by clonal amplification so that the signals generated in the sequencing process become detectable.

3. Sequencing by Synthesis

First, reverse strands are removed before sequencing starts. This ensures that only the forward strands are sequenced and the read out is homogenous and not overlayed with a second sequence which would otherwise render the detected sequence unusable.

A polymerase then "sequences" the insert of the library by adding nucleotides to the complementary strand that are fluorescently labelled. Depending on the chemistry used in the respective machine, either all four nucleotides are labelled differently, or only a subset of the nucleotides contain a label. The label also acts as a terminator, so that the reaction is stopped after the incorporation of one nucleotide. Figure 2 illustrates the process with four fluorescently labeled nucleotides (Courtesy of Illumina, Inc.). After each cycle, the emission of the fluorophore is detected for each cluster using an optical system and thus the identity of the incorporated base is determined. The termination block and the nucleotides of the previous cycle are removed, and the process is repeated until the desired read length is reached. After completion of the sequencing run, the optical signals are translated into the actual sequence, a process termed base calling (Fig. 2).

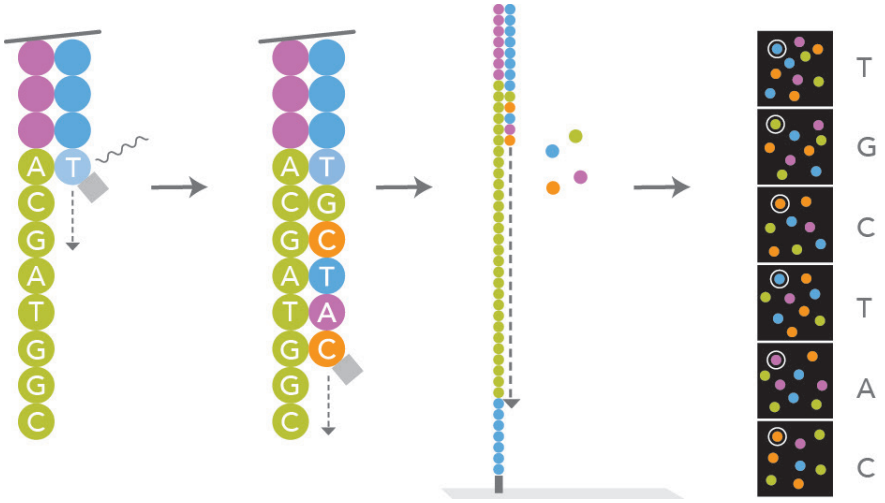


Figure 2 | Sequencing by Synthesis. Sequencing by synthesis technology uses fluorescently labeled A, C, T, and G nucleotides. For each sequencing cycle, a single labeled nucleotide is added to the growing nucleic acid chain. The nucleotide label serves as a terminator for polymerization and after each incorporation the fluorescent dye is imaged. Removal of the dye prior to the subsequent cycle allows incorporation of the next nucleotide. The imaged fluorescent signals are then translated into the nucleic acid sequence. For more information, visit Illumina website.

Image Courtesy of Illumina, Inc.

Following sequencing of the forward strand with the Read 1 primer, the newly created strand is removed. The first index is then read out following the same principle, Index Read 2 is optional. Finally, the reverse strand is read out using the Read 2 primer. To learn more about Sequencing by Synthesis, we recommend the educational resources provided by Illumina.

4. NGS-inherent Sampling Variance

One key facet of sequencing lies in the relationship between the amplification and QC of the libraries, and the sequencing run itself. One way to look at this is by working backwards, starting with the needs of the sequencer – more specifically, the flow cell. For sequencing itself to occur, the lane mix of prepared libraries is strongly diluted following library preparation.

For example, a common scenario would involve diluting a lane mix down to 2 nM, followed by a second dilution by a factor of around 1,000 to a concentration of 2 pM which is used to load on the sequencer. Yet *another* "dilution" occurs on the flow cell itself, where only a subset of the molecules is actually captured on the flow cell.

So, what does this mean for your library preps? It means that the final "yield" of a library prep method is only really important for one thing, and that is quality control (QC). The last step of library generation is PCR, in which the adapter sequences are completed, indices are introduced, and the library is amplified. Since the ultimate output of a library prep method is subsequently diluted to the diminutive amounts required by a flow cell, it is only necessary to generate enough material to be analyzed by quality control instrumentation.

Library quality control quantifies two primary aspects of a library output: concentration and size distribution of each final library. When it comes to library yield, the amount of material needed to reach the detection limit of the QC instruments is much higher than the requirements of a flow cell as described above.

There is an interesting issue that arises when a library is first amplified during PCR and subsequently diluted for sequencing. This "random drawing" effect is the ultimate determinant of which molecules get sequenced from each sample (Fig. 3). There are a few implications of this effect which we will now review.

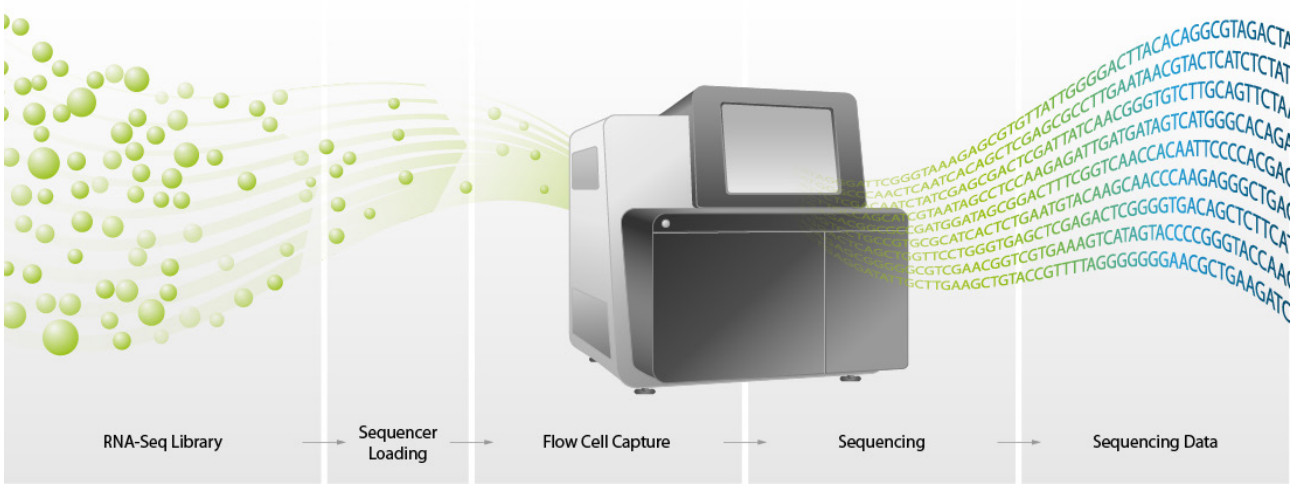


Figure 3 | Only a fraction of the molecules in an NGS-library will be sequenced. *Graphic drawing based on Illumina's NextSeq 500*

5. Why less is more

When one library is sequenced twice on separate runs, it is expected that these runs would yield slightly different results in downstream data analysis. As such, the inclusion of controls in each sequencing run is imperative. The sequencing process itself adds technical noise and has an inherent statistical limit. That is, sequencing data has a ceiling that is reached prematurely, as unavoidable noise prevents an inherently perfect sequencing run.

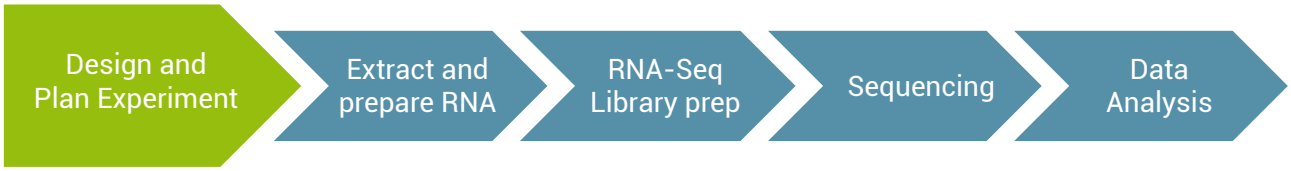
In today's sequencers, data is stored as a FASTQ file. This data is comprised of the collective reads from a sequencing run as well as other statistical measures of the quality of the sequencing run, i.e., the sequencer's own confidence assessment. The factors that impact this quality assessment are a result of the imperfections associated with the sequencing process itself. There are some unavoidable adverse effects which occur regularly, though in small numbers, in any given sequencing run, such as insertions and deletions (INDELS) and base substitutions.

The ideal solution to minimize noise in downstream data analysis is to fit all samples that are to be compared into one sequencing run. Of course, this ideal is rarely possible, and many sequencing projects require multiple runs from which samples are compared. In this case as well, including spike-in controls and reference samples is vital. One step in the pre-sequencing workflow, namely library amplification by PCR, can be a key contributor of avoidable noise. For library amplification PCR, less really is more. The amount of technical noise increases the more a library is amplified. Bear in mind, this library is ultimately diluted significantly before being sequenced. In practice, one obtains better data from libraries that have undergone fewer PCR cycles. This is also why over-cycling can be a serious issue even with Unique Molecular Identifiers (UMIs) present. A future chapter of Lexicon will dive into PCR amplification in greater detail, and we will also introduce UMIs.

What we do here at Lexogen is design our protocols with the highest priority on output integrity. Put another way, Lexogen wants to make sure your data is as good as it can possibly be. One way we do this is by minimizing the disparity between the amount of library required for sequencing and the amount of post-PCR product, thus minimizing the technical noise introduced by the "random drawing" effect. In practice, this means that we are not focused on maximizing the ultimate yield of the library, since doing so would be counter-productive in the pursuit of excellent RNA sequencing performance. Instead, we focus on producing what is ultimately more than enough to sequence, while maximizing data quality.

Experimental and Data Analysis Planning for RNA Sequencing

The path from an intriguing research problem to a well-designed RNA-Seq experiment is an exciting one. There is a wide array of considerations to be made to ensure the success of your project before even starting your experiment. A detailed plan can virtually custom tailor each step in your RNA-Seq experiment to your specific needs and deliver the answers to your wildest research questions.



Thorough planning may be the most important part of an RNA-Seq experiment. When compared to classical, targeted approaches for RNA analysis such as Northern Blotting and RT-qPCR, RNA-Seq experiments can be costly, time-consuming, and fickle. In exchange for these entry barriers, RNA-Seq offers the most comprehensive view of the transcriptome and can be utilized to assess global changes across a multitude of sample types in an unbiased manner.
Understanding your requirements and planning your experiments carefully will increase the likelihood of success and avoid the risk of generating data that fails to answer the fundamental questions of your experiment. With this thorough planning in mind, we have summarized the key considerations for planning RNA-Seq experiments. Stay tuned! The RNA Expertise Hub will offer a dedicated **Quick Checklist for Experimental Planning** coming up as part of our future releases. This checklist and the **"How-to-RNA-Seq" Experimental Planning Guide"** will offer an even more comprehensive collection of methods and parameters that can help you design your RNA experiments.

1. Research Question

The research problem and the sample type make up the core foundation of the experimental plan. It is of utmost importance to determine your specific research question and aims and to design the experiment accordingly.

For example, experiments that are designed to measure quantitative changes in the expression level of genes have different requirements than experiments designed for generating a new annotation for less explored organisms, tissues, or RNA classes. Moreover, it would be very difficult, time consuming, and costly to design just one experiment aimed to satisfy the requirements of both types of experiments.

The primary goal of experimental planning is therefore to determine the main objective and design your RNA-Seq workflow in a way to maximize the output. Below you will find a few considerations to take into account when formulating your aims and objectives.

What kind of data is needed to answer your research question?

Is qualitative or quantitative data needed? For example, do you need information about the properties of transcripts, or do you primarily need to compare their expression level in different samples?

Do you require accurate gene expression data or rather transcript-level information?

Would 3' mRNA-Seq be beneficial, or do you need complete transcript coverage? Do you need information from the whole transcriptome or are you interested in targeting only a subset of transcripts or distinct regions?

What RNA type is of interest?

Would you like to analyze protein-coding mRNA, total RNA incl. non-coding and non-polyadenylated RNAs, small RNA or even all types of RNA?

Are longer sequencing read lengths or even long-read sequencing required?

Transcriptome assemblies and transcript (re-)annotations benefit from longer read lengths. Short read lengths (~75 bp) are the most economical solution for gene expression profiling applications.

How many samples, replicates, and controls are required?

The application itself is an important factor in choosing the right number of replicates, correct sequencing depth, controls, and other sequencing-related parameters, e.g., gene expression profiling experiments benefit from a higher number of replicates (see also requirements for data analysis).

2. Sample Type

The sample type used in an experiment can impact all aspects of the downstream RNA-Seq experiment. The sample itself affects the choice of RNA extraction, suitable pre-treatment, the number of controls and replicates required to answer the research question with confidence, and the choice of the library preparation kit itself.

What is your species / organism of interest?

Is the genome annotated, are polyA-tails present, are small RNAs described?

Is the sample type highly heterogenous?

More replicates may be required to account for variance and spike-in controls can help to assess the sequencing data with confidence.

Is the sample / RNA degraded?

For example, an appropriate RNA extraction method for RNA of all sizes should be chosen, long-read sequencing is not applicable for degraded RNA and ribosomal RNA depletion should be chosen over poly(A) selection for whole transcriptome sequencing.

How much material is available?

Limited complexity and low input samples have lower read depth requirements, for example, a library derived from 1 ng RNA input will have lower complexity than a library derived from 100 ng input RNA and does not require as much sequencing depth.

3. Transcriptome Complexity

Organisms with lower transcriptional diversity may not require as much read depth for sufficient transcriptome coverage and thus more libraries can be multiplexed, e.g., bacterial transcriptomes are much smaller than mammalian transcriptomes, less genes are expressed at the time of sampling and often less transcriptional isoforms are present per gene.

Sample complexity is not only a consequence of the complexity of transcriptomes between different species or domains of life. The transcriptomic landscape and complexity can also vary between different tissues, biofluids or cell types within the same organism due to RNA content, expression patterns, over-abundant tissue-specific transcripts, etc. Tailoring your workflow and sequencing strategy to your specific sample type can optimize the data quality and save time and overall costs.

4. Choosing the Right Library Preparation Method

Designing an RNA-Seq workflow for differential expression analysis from whole blood RNA samples. Curious to see what we are planning?
Click here to find out more.

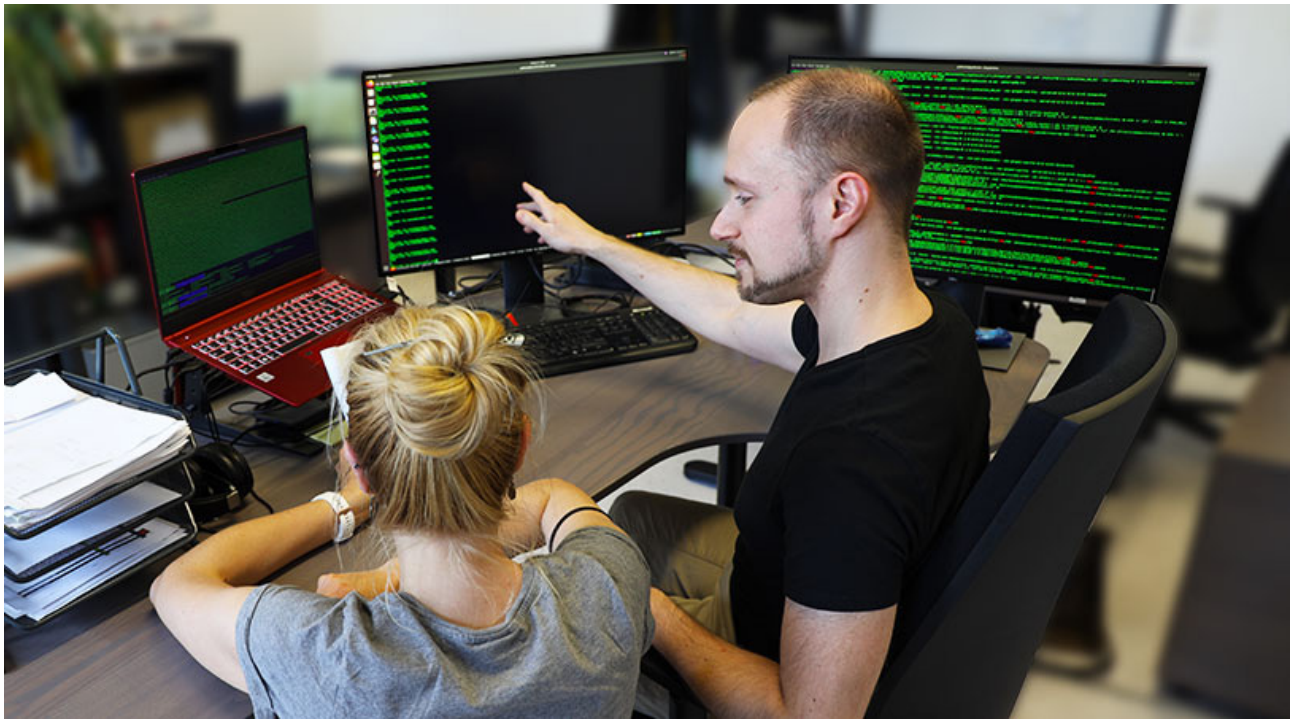
The choice of library preparation impacts the kind of data that can be obtained in the RNA-Seq experiment and ultimately needs to be aligned with the level of information that is required. Apart from sample type-related constrains, e.g., poly(A) selection is not possible for organisms that do not contain poly(A) tails or when working with degraded samples, the library preparation method needs to be chosen according to the research question.

For example, a 3′ mRNA-Seq approach generates sequencing reads localized to the 3′ end of mRNAs. Therefore, it is a highly convenient method for multiplexing a large number of samples, does not require poly(A) enrichment prior to library preparation, and allows to accurately quantify gene expression with minimal computational resources. You can find short overview of read depth requirements in our blog article "How many reads do I need for my RNA-Seq samples?"

These features make 3′ mRNA-Seq the method of choice for gene expression profiling, especially for high-throughput projects. However, due to the fact that reads are localized to the 3′ ends of the transcripts, this method is not suitable to assess alternative splicing within the transcript body, investigate differential transcript usage, or for the identification of transcript isoforms. Thus, if this level of information is required, a whole transcriptome library preparation method that provides full transcript coverage would be the appropriate choice.

In contrast to 3′ mRNA-Seq preps, whole transcriptome library preps usually require either poly(A) enrichment or rRNA depletion to focus the reads on the transcripts of interest. Which enrichment / depletion strategy is chosen depends on the sample itself as well as on the nature of the RNA of interest. If you are interested in polyadenylated mRNAs only, poly(A) enrichment is the method of choice. However, if you want to analyze also non-coding RNAs which may lack poly(A) tails, depletion of ribosomal RNA should be used instead. And if you are interested in small RNAs, you need to use an extraction procedure and library preparation method that is suitable for these short transcripts.

5. Data Analysis Planning



Once you have set your mind on a suitable library prep method, the parameters for data analysis must be taken into consideration. Ideally, the optimal parameters for data analysis, e.g., the minimum number of replicates required for a statistically sound analysis, are best discussed with a Bioinformatician prior to starting your experiment. This part of the planning process ensures that the generated data set fulfills the requirements to run the data analysis pipelines needed to answer your research question.

Below you will find a few questions that should be assessed during data analysis planning.

Is the organism characterized? Is an annotation available for the specific research question? Or does it need to be built or refined?

Are other experiments needed to provide the basis for your research? For example, do you need to incorporate other data, e.g., long-read sequencing and transcript assemblies or annotation of 3′ UTRs before ultimately being able to assess your research question on your organism of interest?

What kind of analysis is required?

For example, gene expression profiling and differential expression analysis use different tools than transcriptome assemblies and have other requirements, i.e., more replicates at lower sequencing depth vs. less replicates at much higher sequencing depth

Data evaluation and statistics

How many replicates are needed to answer the question with high confidence and statistical significance

How many sequencing reads are needed?

How much sequencing depth is needed for the particular application you are interested in?

Is it possible to provide the required read depth and optimal number of replicates?

Should you sequence less replicates deeper or more replicates at reduced depth? Often a compromise needs to be found between the number of replicates and the sequencing depth. Staged sequencing can also be used as a compromise, i.e., sequencing the same samples in multiple runs if needed and adding up the read depth can provide higher depth for a larger number of replicates.

What are the specific and agnostic controls that are needed?

Which controls are needed for optimal data analysis, e.g., sample controls, application-specific controls and general sequencing controls to assess the performance of the workflow and data analysis, e.g., ERCCs and SIRVs.

Are there specific requirements met for the tools I want to use?

Some tools require replicate data as input for analysis.

6. Checking Your Experimental Design Using a Pilot Experiment

After carefully planning the experiment, it is time to prepare the samples for sequencing. Before starting a large-scale or long-term experiment, it is extremely useful to conduct a pilot experiment with a representative but smaller set of samples to check if the chosen experimental parameters deliver the required results and the data analysis requirements are met. In case you have various options and methods to choose from, a comparison can be included allowing you to evaluate the different workflows in parallel and pick the best option for your needs.

After assessing the results from the pilot, you can still make adjustments to the experimental setup and parameters before diving into a larger experiment.

The following chapters will now take you to the lab, shed light on sample handling, give advice for best practice handling with a special focus on difficult sample types, and introduce further useful tools and quality measures to help you generate the best RNA-Seq data possible to advance your research.

RNA Extraction and Quality Control

In this chapter we will finally move to the lab and discuss the key steps in RNA extraction and considerations to take when choosing how you will isolate, extract, and store your RNA prior to RNA sequencing.

First, you need to determine your overall goal(s) and plan out the downstream steps prior to choosing an RNA extraction method.

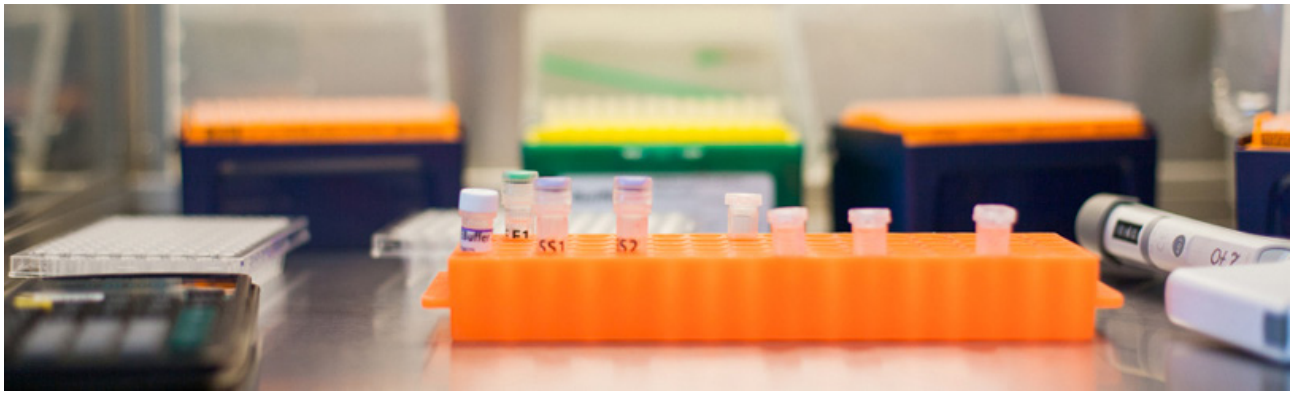
- ✓ Are you interested in analyzing mRNA only, total RNA, or a specific subset of RNAs (i.e., small RNAs such as microRNAs)?
- ✓ Will you be working with samples that allow isolation of high-quality RNA or with challenging sample types that are prone to RNA-degradation or fixed material?
- ✓ What type of enrichment, if any, is desired (i.e., ribosomal RNA-depleted RNA, poly(A) RNA, etc.)?

The answers to all of these questions will help determine what type of RNA extraction method to utilize.

Check out our third Chapter focusing on Experimental and Data Analysis Planning and our Checklists to map out your workflow prior to beginning any wet lab procedures. The table below contains useful tips on what to consider when choosing an RNA extraction method.

Research Interest	Parameter	Considerations
Are you interested in analyzing mRNA, total RNA, or a specific subset?	mRNA and total RNA	The average length of mammalian mRNA is ~2200 nucleotides. In general, RNA molecules can range between 15 and 17 000 nucleotides.
	Small RNAs only	Small RNAs can be as short as ~15 nucleotides and typically range between 20 – 30 nucleotides for eukaryotes. In bacteria, small RNAs can be up to ~300 nucleotides. Make sure your extraction method is suitable for small RNA.
	Membrane-associated mRNAs	Basic hot phenol extraction can yield more membrane-associated mRNAs as the procedure facilitates the membrane detachment [1].
Will you be working with high quality, intact samples or low quality, degraded samples?	High quality (e.g., bacterial cultures, tissue / cell cultures, freshly frozen material)	Some sample types might need specialized disruption, e.g., bacteria / plant require mechanical or enzymatic disruption of cell wall.
	Low quality (e.g., formalin-fixed paraffin-embedded [FFPE] samples, biopsy samples)	Challenging and fixed sample types contain shorter, non-intact transcripts. Ensure the extraction method does not eliminate shorter fragments. Extract from a higher input amount for FFPE samples (some RNA molecules can be crosslinked and this inaccessible during library preparation).
What type of enrichment, if any, will be needed?	No enrichment (total RNA)	Ensure the RNA extraction method isolates all size transcripts.
	Ribosomal RNA-depletion	Ribosomal RNA depletion can also be used for low quality RNA samples and is advised for best practice when the isolated RNA has a RIN / RQN value < 8.
	Poly(A) enrichment	Ensure that the resulting RNA is high quality (RIN / RQN > 8) to use poly(A) enrichment.
What type of sequencing will be used?	Short read sequencing (50 – 500bp)	Short read sequencing is also suitable for degraded samples.
	Long read sequencing (>500bp)	Requires intact RNAs, mild lysis and RNA extraction conditions should be chosen to preserve full length RNA molecules.

Table 1 | Key considerations for RNA extraction based on research interest



RNA is a highly susceptible molecule and degrading RNases are ubiquitous. Therefore, special care should be taken to inactivate RNases during the procedure and avoid contamination of the extracted RNA. Harvesting the samples by flash / snap freezing in liquid nitrogen or on dry ice preserves the RNA integrity by terminating all biological processes in the cell, including RNA turnover by endogenous RNases. Alternatively, RNA stabilizing agents can be added to the sample to safeguard the RNA prior to extraction. These chemical solutions permeate fresh (non-frozen) tissues, cells, and liquids and inactivate RNases in the sample and allow storage even at room temperature or in a refrigerator before RNA extraction.

There are numerous methods available for extraction RNA from your sample, but the basic steps remain fairly conserved:

1 Cell disruption:

Disruption of tissues and lysis of cells is a critical step in the RNA isolation process as it impacts both the quality and quantity of isolated RNA. The commonly used forms of cell lysis are either chemical, mechanical, or enzymatic lysis. The choice between these two possibilities will ultimately depend on the sample you are working with. In the end, a homogenous solution with no visible particulates should be obtained.

2 Removal of DNA and proteins:

Ensuring no contamination is carried over into your final RNA sample is important, and this is typically achieved by a single-step technique using an acidic solution consisting of guanidinium salts, acetate, phenol, and chloroform. Guanidinium salts are chaotropic chemical compounds that denature proteins including Nucleases and thus have a protective function in nucleic acid isolation procedures. Phenol and chloroform are organic solvents that trap and separate proteins and lipids while nucleic acids stay in the aqueous phase. Acetate is used to generate acidic conditions. At pH 4-6 also genomic DNA (gDNA) will be trapped in the organic phase while the RNA will still be retained in the aqueous phase. This procedure can thus minimize the carry-over of gDNA for a highly pure RNA preparation.

Phenol-free and column-based RNA extraction methods also use denaturing agent but rely on binding nucleic acids to silica-based columns instead. These methods are often used by researchers as they offer a fast and highly convenient extraction, however, additional DNase treatment steps are required for challenging applications such as RNA-seq as the gDNA is usually not removed as efficiently.

3 Denaturation and inactivation of RNases:

Due to the unstable nature of RNA and ubiquitous presence of RNases, strong denaturants and chaotropic agents (such as guanidinium salts and phenol) are typically utilized in the early steps of RNA isolation to inhibit endogenous RNases.

Successful extraction of RNA relies on good laboratory practices and RNase-free environments. Gloves should be worn at any time and speaking over opened tubes should be avoided.

4 Precipitation:

RNA is concentrated and further purified by precipitation with ethanol or isopropanol. Alternatively, RNA can be bound to a silica column. In presence of ethanol, nucleic acids stay bound on the column and can thus be washed and freed from contaminants. In absence of ethanol, nucleic acids are eluted.

When minute amounts of RNA need to be purified, precipitation is the method of choice as losses are reduced compared to column-based methods. Further, the use of carriers, such as glycogen increases the recovery of nucleic acids and facilitates handling of the pellet formed upon precipitation.

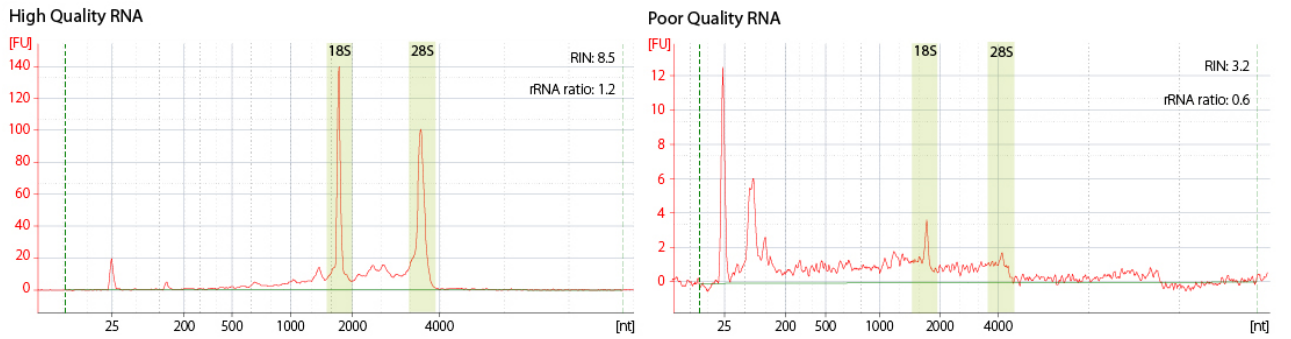


Figure 1 | Bioanalyzer traces of extracted RNA run on an Agilent RNA Pico Chip. RNA was loaded and analyzed according to manufacturer's instructions, the RIN values and rRNA ratios are shown for high quality and poor quality RNA.

5 Quality Control:

Following extraction, the RNA should be quality controlled to ensure purity and integrity. RNA preparations should be free of salts, metal ions, and organic solvents, or other contaminants, which can be carried over from RNA extraction. Several sources of contamination can be detected with a UV-Vis spectrophotometer. An acceptably pure RNA sample should have an A260/A280 ratio between 1.8 and 2.1. The A260/A230 ratio should be approximately 2. Several common contaminants including proteins, chaotropic salts, and phenol absorb strongly between 220 and 230 nm and can often be identified as peaks in this region. Contamination with any of these generates a lower A260/230 ratio. As these contaminants can have a negative impact on the downstream processing steps in an RNA-seq workflow, they should ideally be removed by an additional purification step, such as precipitation for best quality results.

The integrity of the RNA can be assessed by a variety of methods. Microfluidic assays, such as Bioanalyzer or Fragment Analyzer RNA assays, are most commonly used. Based on the ratio of the rRNA peaks the RNA quality score (RIN or RQN) is determined. Figure 1 shows exemplary traces from a microfluidic RNA assay for high and low quality RNA on Bioanalyzer.

When assessing your extracted RNA, choose the kits suitable for the desired size range and the expected RNA amount and stay within the specified detection range.

6 Storage:

After extraction, the RNA is ideally stored in a buffer at pH 7 at -20 °C or -80 °C and freeze / thaw cycles should be kept at a minimum, e.g., by storing the RNA in various aliquots. Additionally, reducing agents, such as DTT, chelators, such as EDTA, and RNase inhibitors can be added for long term storage. As these compounds interfere with optical density measurements, these additives are best supplied after the concentration measurement and should be added to the buffer for the blank when remeasuring.

The extraction method chosen will strongly depend on your sample type and overall goal, so be sure to read up on which extraction method is right for you.

Sample type-specific considerations

Find out what you need to consider when you are extracting RNA from difficult sample types, such as plants, blood, or FFPE material.

Plants

Plants can be considerably challenging due to the presence of secondary metabolites, polyphenols, and polysaccharides. Based on physical and chemical properties of these inhibitors, which are similar to nucleics acids, these inhibitors can coprecipitate with RNA irreversibly. Due to the negative influence they can have on downstream steps such a library preparation, care should be taken to remove them from the RNA. Dedicated kits for RNA extraction from plants and phenol/chloroform-based approaches are especially useful to remove these inhibitors.

Formalin-fixed paraffin-embedded (FFPE) tissue

FFPE samples also pose several challenges to RNA Extraction: cross-links are introduced between macromolecules during fixation, the RNA is often highly degraded, and gDNA contamination is common during extraction. Therefore, the appropriate RNA extraction kit should be used for FFPE samples, e.g., a kit that is suitable for fragmented RNA and minimizes the carry-over of gDNA such as Lexogen's SPLIT RNA Extraction Kit. It would also be beneficial to use a kit specifically designed for FFPE samples, where deparaffinization and treatment with Proteinase K are essential parts of the workflow.

As FFPE samples are commonly used to store precious clinical and biobank samples, solutions to work with this sample type are in high demand. We will return to this topic in several of our later chapters and will also publish a dedicated Handling Guide for FFPE Samples, so stay tuned to find out more about these exciting samples.

Blood

When working with blood-derived samples, a sample-specific RNA extraction method should be chosen. RNA can be extracted from whole blood or the blood can be preprocessed to isolate serum, plasma, or buffy coat peripheral blood mononuclear cells (PBMCs). All of these sample types have their specific requirements, e.g., serum and plasma often contain only low amounts of RNA which can be fragmented, protein-bound or encapsulated in extracellular vesicles. Therefore, specialized kits suitable for low RNA quantities and short fragments should be used. Blood cells in general contain more DNA than RNA, so that in contrast to other eukaryotic cells, gDNA carry-over in RNA preparations is very common. Additionally, the high amount of DNA and proteins can lead to highly viscous lysates and therefore, the input amount should be reduced. Also, globin mRNA which encodes for subunits of hemoglobin is highly abundant in blood samples and can amount to 30-80 % of reads in an RNA-Seq experiment. To effectively remove globin already during the prep, red blood cell lysis is often employed, e.g., as present in Lexogen's SPLIT RNA Extraction Kit for Blood. For best results, freezing should be avoided. Especially when red blood cell lysis should be used to remove globin mRNA, using fresh, non-frozen blood is mandatory. RNA stabilizing agents can be used to preserve the blood RNA, but these may interfere with the lysis. In this case, a different globin depletion strategy should be used.

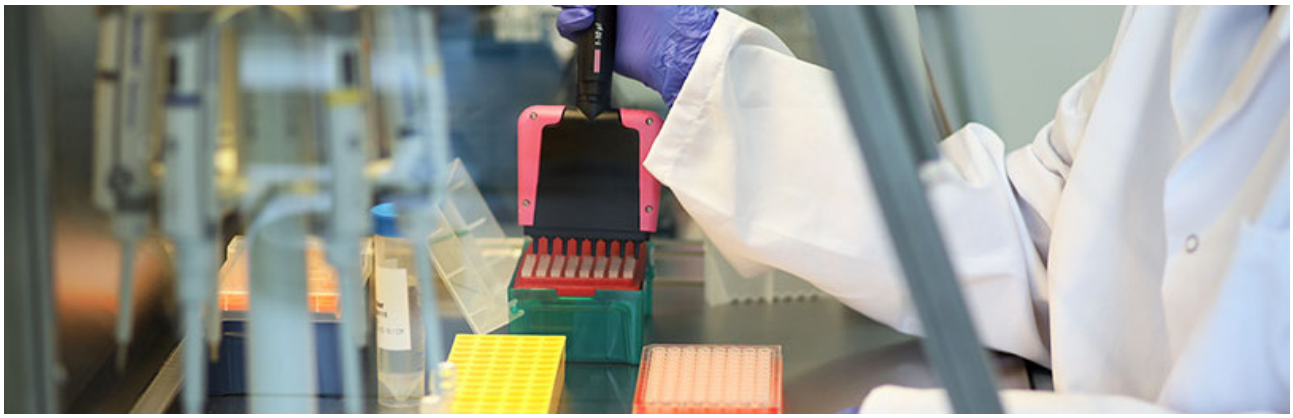
Blood is a very challenging sample type that requires special consideration in various steps of an RNA-seq workflow, we will return to this topic also in our upcoming chapters discussing DNase treatment and in our Practical Tips and Handling Guide for Blood Samples.

Literature:

Scholes, A.N., Lewis, J.A. Comparison of RNA isolation methods on RNA-Seq: implications for differential expression and meta-analyses. *BMC Genomics* 21, 249 (2020). DOI: 10.1186/s12864-020-6673-2

DNase: To Treat or Not to Treat

In an RNA sequencing experiment, the quality of the RNA used for library generation is of the utmost importance. Put another way, high quality RNA will have the best chance of leaving the researcher with high quality data. There are many instances where researchers may find themselves in a place where high quality, high purity RNA starting material is impossible, such as working with FFPE samples (this will be covered in a future Lexicon chapter). Thus, we can think of this through yet another lens: *getting the most from what you have*. Whether you are working with FFPE samples or cell lines, the goal is to get the most from what your starting material can offer. In many cases, DNase treatment is a key step in maximizing the quality of the RNA samples used for sequencing.



With few exceptions, any RNA sample will contain some amount of contaminating genomic DNA (gDNA). It is not surprising then, that many RNA extraction kits recommend, or even require DNase treatment of RNA samples before proceeding to challenging downstream applications, such as transcriptome analysis.

In RNA-Seq applications, random primers or short oligo(dT) primers cannot distinguish between RNA and DNA, and will also hybridize to residual gDNA. In addition, reverse transcriptases are promiscuous enzymes which are able to use DNA as template molecule. As a result, unwanted gDNA will be channeled through the entire RNA-Seq workflow, which can cause biases and quantification issues during the final data analysis steps. Therefore, it is critical to remove any residual gDNA to obtain the best quality data.

1. Genomic DNA Removal Methods

The most common means of DNA removal is by DNase digestion. DNase, short for Desoxyribonuclease, is a DNA-specific endonuclease that cleaves single- and double-stranded DNA, leaving behind 5' phosphorylated oligonucleotide products. Because of this versatility it is used in a wide range of biological applications. It is important to note that DNase should be removed afterwards as residual DNase may also affect downstream reactions in the library preparation.

1) On-column DNase Treatment (during Extraction)

On-column digestions are commonly used during the RNA extraction procedure. In these methods, the lysate is loaded onto a column where a filter substrate binds the RNA and contaminant DNA. A series of wash steps and an incubation step with a DNase containing solution are used to digest genomic DNA leaving behind your intact RNA. The RNA is then washed before it is finally eluted off the column.

Column digestions are among the most commonly used methods, though they are not without their pitfalls. The two main drawbacks – on-column digestion increases the time requirement of the RNA extraction and can decrease the RNA quality in case contaminants are not efficiently removed prior to the DNase treatment step. Performing the treatment on column-bound DNA can also result in residual gDNA carry-over. Despite these downsides, column digestions are frequently performed due to their ease of use and accessibility.

2) Acid Phenol Extraction of RNA

Acidic buffered phenol / chloroform extraction is an established method used commonly to minimize carry-over of gDNA. It is used in labs all around the world as it yields high-quality, high-purity RNA. For many samples, no additional DNA digestion is needed after acid phenol extraction. For more details see our previous chapter on RNA Extraction.

3) DNase I Digest after RNA Extraction

DNase treatment can also be done in solution following RNA extraction. As opposed to the column-based method described above, DNase I digest in solution is commonly accepted as a more efficient and thorough way to eliminate gDNA from an RNA sample. Routinely, extracted RNA is mixed with DNase and reaction buffer and incubated either at room temperature or 37 °C for 15 min to 1 hour. Following incubation, the DNase is inactivated or removed by one of the clean-up methods described below to stop the reaction and purify the RNA for further processing.

2. DNase Clean-up Methods

Remaining DNase should be removed from the sample before the experiment proceeds. If DNase is carried over into library preparation, primers initiating reverse transcription may be degraded ultimately affecting the efficiency of the library generation. The method of DNase clean-up is perhaps just as important for the sample quality as the use of DNase in the first place. To minimize DNase presence in the final RNA sample there are a number of options available, each with their own strengths and weaknesses.

1) Purification

Column- or bead-based purifications and ethanol precipitation are widely used to remove not only DNase itself but also the reaction components. Using a clean-up leaves you with the pure RNA sample eluted in water or buffered TRIS solution. This allows further processing of the sample without any interference from the reaction components used in the previous DNase digest.

Column-based clean-up is quick and easy method to purify nucleic acids after enzymatic reactions. The reaction is mixed with binding buffer and directly loaded to the column. The intact RNA is bound to the filter substrate while proteins, i.e., DNase, short DNA oligonucleotides and residual buffer components are washed away before the RNA is finally eluted.

When working with many samples, bead-based purification can be used to increase the number of samples that can be handled at a time or even automate the clean-up. In this method, nucleic acids are bound to the surface of specific magnetic beads, while short fragments, proteins and buffer components remain in the supernatant and are removed. The beads can be collected using magnets, and are washed before the nucleic acids, in this case the cleaned-up RNA, is eluted.

Ethanol precipitation is another common clean-up method: The reaction mix is precipitated in presence of ethanol and salt (commonly LiCl is used for RNA precipitation) and if needed a carrier is added, e.g., glycogen. The pellet is washed to remove reaction components before it is resuspended in buffer or water.

Ethanol precipitation is often used for precious samples as it preserves the valuable sample by minimizing material loss.

Even though purification methods are the clear choice for obtaining the cleanest possible sample, other methods of DNase inactivation are often preferred due to convenience, cost- and time-savings.

2) Heat inactivation

Using a brief incubation at elevated temperatures is yet another popular method of inactivating DNase. While heat inactivation does not remove reaction components, the simplicity of the method is its major benefit, requiring only 5 minutes at ~75 °C. However, due to the temperature and buffer conditions in the reaction, the RNA can also be fragmented easily. As RNA integrity is of utmost importance for RNA-Seq, heat inactivation should be avoided especially when working with lower quality material.

3) EDTA Chelation

The RNA fragmentation associated with heat inactivation as described above can be mitigated by the addition of EDTA. While this sounds good at surface value, special care should be taken with the amount of EDTA added to the reaction. Too much EDTA can chelate the divalent metal ions required for enzyme activity in reverse transcription, a crucial step in RNA-Seq library preparation.

4) Proteinase K

An additional method for DNase inactivation is Proteinase K treatment. While Proteinase K is particularly effective at digesting proteins and thus deactivating DNase, Proteinase K itself needs to be removed to ensure that the enzymes required for subsequent reactions are not inactivated by Proteinase K carry-over.

3. Detecting gDNA in your Sample

When is DNase treatment required for your RNA sample and how can you decide if the gDNA really was efficiently removed? As you can imagine, due to the similarities between these nucleic acids, detecting gDNA in an RNA sample is difficult. But there are a few methods that can be used to determine whether or not gDNA is present:

1. Optical Density (OD) measurements using a spectrophotometer and comparing emissions at different wave lengths can give you an indication of the purity of your RNA sample, e.g., an OD 260/280 value below 2 can indicate DNA contamination.

2. Some spectrophotometric assays using fluorescent dyes specific for either RNA or DNA can distinguish between RNA and DNA within a sample.

3. Agarose gel electrophoresis can also show gDNA in the high molecular weight region of the gel (Fig. 1A).

4. Running the extracted RNA on a Fragment Analyzer using extended run time settings can reveal gDNA contamination in a similar way to that of the gel above: a high molecular weight "bump" in the trace indicates the presence of gDNA in the sample (Fig. 1B).

Most of these methods only detect rather high levels of gDNA presence in a sample. Unfortunately, RNA-Seq is so sensitive to gDNA contamination that the amount required to negatively impact the process falls below the detection threshold of some of the methods described above. This is a primary contributing factor to the popularity of DNase treatments in RNA-Seq workflows.

The best approach to detect residual gDNA and ensure the RNA preparation is indeed fully DNA-free is to run a PCR or qPCR with primer pairs for a specific set of marker genes. Commonly referred to as house-keeping genes, there are some popular genes such as GAPDH, genes encoding RNA-polymerase subunits, actin, rDNA loci, and others. As PCR can detect trace amounts of DNA from as low as one molecule, PCR is the method of choice to exclude gDNA contamination in workflows that require highly pure RNA that needs to be absolutely free of DNA.

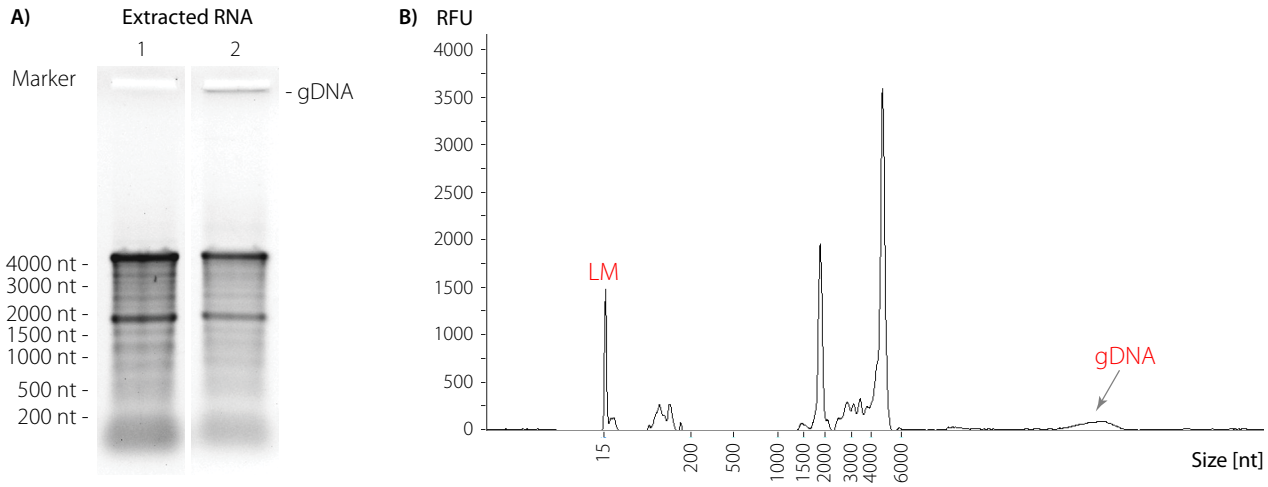


Figure 1 | Assessing your RNA sample for genomic DNA (gDNA). A) Stained agarose gel assessing RNA extracted using an RNA extraction method with (lane 1) and without gDNA removal (lane 2). A high molecular weight band is visible in lane 2, indicating the presence of genomic DNA. B) Fragment Analyzer trace with extended runtime assessing RNA extracted from white blood cells. A high molecular weight "bump" indicates gDNA contamination of the sample.

4. To Treat or Not to Treat

The question remains, when is it necessary to treat the RNA preparation with DNase, and when is it unnecessary? In general, DNase treatment should always be considered for the following circumstances:

- ✓ When working with blood, DNase treatment is essential. Unlike other eukaryotic cells, e.g., cell lines, blood cells contain more DNA than RNA. This makes gDNA carry-over highly likely, even when using RNA extraction methods such as acidic phenol / chloroform extraction.
- ✓ Degraded RNA samples are very likely to contain fragmented pieces of DNA with lower molecular weight than that of intact gDNA. These short DNA sequences can be co-isolated with RNA and thus a DNase treatment is recommended.
- ✓ In FFPE samples, as degradation increases, so does the occurrence of cross-links between macromolecules and thus DNA carry-over. A DNase treatment is recommended.
- ✓ Mechanical disruption is very harsh on samples and commonly, gDNA carry-over is observed due to sheering / fragmentation of the gDNA in the process. Here, we also recommend DNase treatment.
- ✓ Finally, sample types with large quantities of short, extra-chromosomal DNA should be treated with DNase. This includes bacteria, which can carry plasmids that reach high copy numbers. These copies increase the DNA content and shift the DNA:RNA ratio more heavily in favor of DNA, making very difficult to avoid carry-over during RNA extraction. DNase treatment is therefore also recommended for bacterial samples.

When can DNase treatment be omitted?

As you have seen, some RNA extraction procedures are more suitable to minimize gDNA carry-over than others, especially acidic phenol / chloroform extraction can remove gDNA from many standard sample types. Also, 3' mRNA-Seq library preps tend to pick up less gDNA background than random primed whole transcriptome library preps, due to the fact that 3'-Seq methods rely on poly(A) stretches for priming. When both are used in combination and the RNA is of high quality and derived from an unproblematic sample type, the risk of compromising your RNA-Seq experiment with DNA contamination is minimal.

Those running targeted sequencing experiments can breathe a sigh of relief, as it were. DNase treatment is not needed for most RNA-Seq applications using targeted primers for specific genes of interest. The risk of contamination with accessible gDNA for exactly this location is low and genomic loci for primer binding are often far from the respective regions of the transcript (e.g., due to the presence of introns), and thus DNase treatment is not required.

RNA Pretreatment: Enrichment or Depletion?

After extracting RNA from your samples, removing residual DNA, and checking the quality of your RNA preparations, you need to decide if and how to enrich the RNA population of interest. Total RNA is comprised of large amounts of ribosomal RNA (rRNA) which can make up between ~80 – 98 % of all RNA molecules in a sample. For most RNA-Seq applications, the removal of rRNA or the enrichment of polyadenylated transcripts is required to focus the sequencing capacity on the desired parts of the transcriptome and to economize the experiment.

Apart from ribosomal RNA, samples can contain additional abundant transcripts, e.g., globin mRNA in blood samples can account for ~30 – 80 % of all mRNA molecules. Without the removal of abundant transcripts, the majority of reads obtained from an RNA-Seq experiment would be derived from these undesired RNA species taking up valuable sequencing space and limiting the multiplexing capacities within the experiment. Several methods can be used for enrichment of desired RNA, or depletion of undesired RNA which we will discuss in this chapter.

1. Poly(A) Enrichment

The most common pre-processing method used in RNA-Seq is poly(A) selection. Poly(A) selection is used to "fish" polyadenylated RNA species from a total RNA solution. Thereby mature, coding mRNAs are enriched providing the basis for mRNA-Seq workflows. During this procedure, the RNA is first denatured to remove secondary structures and make the poly(A) tails accessible. Oligo(dT) stretches attached to a solid surface, most often to magnetic beads, are then hybridized to the poly(A)-containing RNA molecules. Following hybridization, the supernatant consisting of non-polyadenylated species is removed. The beads are washed prior to elution of the poly(A)-selected RNA. Elution is achieved by elevating the temperature, so that the poly(A) – poly(T) base pairing interaction is resolved, and the selected RNA is eluted from the beads into the solution (Fig. 1).

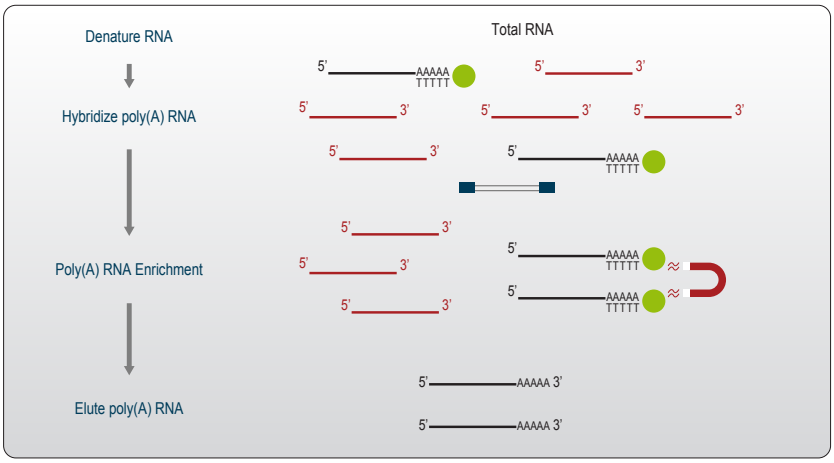


Figure 1 | Schematic overview of poly(A) RNA selection. The workflow is adapted from Lexogen's Poly(A) Selection Kit.

The selection process excludes the majority of rRNA molecules as these are not polyadenylated. Only ~2 – 5 % of reads are commonly mapped to mitochondrial rRNA, which carries a poly(A) tail and is co-purified.

Poly(A) enrichment is a very cost-efficient and fast pre-processing method that allows selection of mainly protein-coding mRNA. It can be used for all species that possess poly(A)-tailed RNA to remove undesired rRNA and concentrate sequencing reads on mRNA.

However, there are two main drawbacks of poly(A) enrichment:

1 First, it can only be used for species that possess poly(A)-tailed RNAs. Therefore, it is restricted to eukaryotes and cannot be used for prokaryotes. However, even eukaryotes possess transcripts of interest that lack poly(A) tails. They will be removed together with rRNA during poly(A) selection. These transcripts encompass microRNAs, small nucleolar RNAs (snoRNAs), transfer RNAs (tRNAs), some long non-coding RNAs (lncRNAs), and even protein-coding mRNAs such as histone mRNAs. As a result, researchers interested in those types of RNA or prokaryotic species commonly utilize rRNA depletion instead of poly(A) selection.

2 Second, poly(A) enrichment requires high quality RNA (RIN / RQN > 8). Degradation leads to breaks within the transcript body and due to the selection of the poly(A) tail, the 3' ends are enriched while the more 5' sequences would not be captured, leading to a strong 3' bias for degraded RNA inputs. Therefore, ribo-depletion or 3' mRNA-Seq are the methods of choice for working with degraded RNA.

2. Poly(A) Enrichment During Library Preparation

It is also possible to select for polyadenylated transcripts during library preparation by using oligo(dT) priming during reverse transcription. In this case, cDNA is generated primarily starting from the 3'UTR right at the beginning of the poly(A) tail of mRNAs. This eliminates the requirement for poly(A) selection by the magnetic bead-based approach described above. Therefore, this principle is commonly used in 3'mRNA-Seq protocols, such as QuantSeq. The complete workflow for 3'-Seq is efficiently shortened and due to the focus on 3'ends it is also suitable for use with degraded RNA.

It is also possible to generate full-length cDNA by oligo(dT) priming. For these protocols, the reverse transcription reaction is optimized for the generation of long fragments and the 5'end is routinely enriched by cap-dependent capture methods or template switching.

3. Enzymatic Removal of Abundant Transcripts – Duplex-specific Nuclease (DSN) Treatment

Duplex-specific Nuclease (DSN) is a thermostable nuclease isolated from Kamchatka crab. The enzyme possesses a strong affinity for double-stranded DNA (dsDNA) which is cleaved efficiently while enzyme activity towards single-stranded DNA (ssDNA) is limited. DSN has been used in life science fields for DNA copy number normalization and is not only used in Next Generation Sequencing (NGS) approaches [1,2] but also in forensic analysis of low copy number DNA.

Different DNA copy numbers in an RNA-Seq experiment are the result of two main factors.

1 The expression level of transcripts in a cell varies between several orders of magnitude: while some transcripts can be present with more than 10,000 copies per cell, others may only be expressed at a very low level with only 1 – 2 copies.

The most abundant transcripts in total RNA samples are rRNA, tRNA, and housekeeping mRNAs. Also, tissue- or sample-specific overabundant transcripts fall into this category.

2 PCR and amplification bias can lead to preferential amplification of certain molecules while others are under-represented and can thus also be a source of copy number variation.

How Are Abundant Sequences removed by DSN Treatment?

In RNA-Seq approaches, DSN treatment is used to partially normalize the concentration of cDNAs that reflect the dynamic range of the transcripts. This is achieved by removal of abundant transcripts. DSN treatment is commonly performed after cDNA first and second strand syntheses, however, it is also possible to use DSN after first strand synthesis when the RNA template is not yet removed.

The DSN reaction then uses the hybridization properties of the newly synthesized cDNA molecules (Fig. 2). Following cDNA synthesis, the molecules are denatured by a brief incubation at a high temperature before the reaction is cooled again. Upon decreasing the temperature, the complementary cDNA strands are re-annealed (a process called renaturation). Due to the higher concentration and thus higher chances to interact with the complementary strand, abundant cDNA strands re-anneal faster and more efficiently than lower abundant cDNAs. Therefore, the majority of double-stranded cDNA will be derived from abundant transcripts, while the cDNA derived from the medium and lowly expressed transcripts will stay single-stranded. The double-stranded cDNA fraction is then cleaved by DSN whereby all abundant sequences are removed and the molecules in the pool are normalized to a similar concentration level. Finally, the remaining cDNA molecules are amplified in a PCR reaction to generate sequencing-ready libraries (Fig. 2).

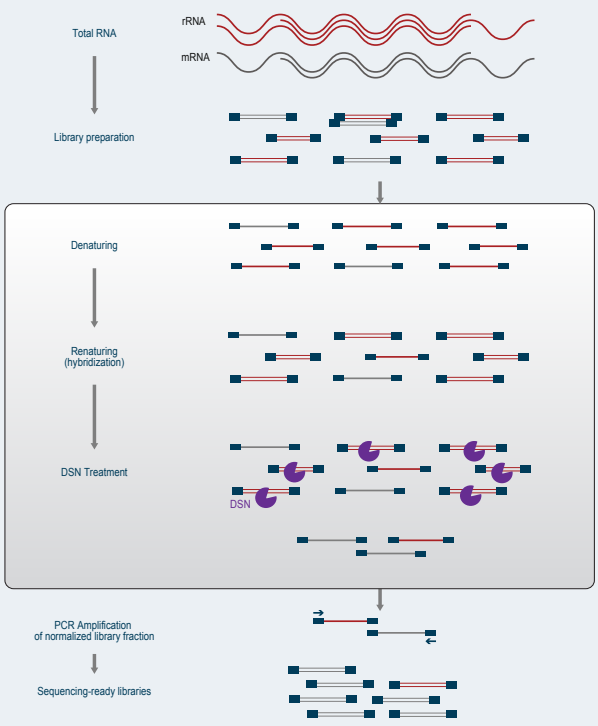


Figure 2 | Enzymatic removal of abundant sequences using Duplex-specific Nuclease (DSN) treatment.

DSN treatment is routinely used in a wide range of applications, but especially for annotation-based approaches as it is quite universal and has the potential to deplete any abundant sequence. It is useful for obtaining transcriptome information from less characterized species for which targeted depletion methods using specific probes are not available, as well as for species that do not possess poly(A) tails and therefore cannot be enriched for mRNA by poly(A) selection [3]. Some RNA-Seq kits and universal depletion methods therefore use DSN treatment in their workflows.

The major drawback of this method is that the depletion is unspecific and targeted towards any abundant sequence. If your transcript of interest belongs to the higher-copy number group, it might also be subject to DSN-mediated degradation. In addition, special care should be taken when quantitative information is required to evaluate changes in expression levels. Depending on the input used, DSN treatment may normalize the concentrations in a way that makes it impossible to quantify the changes in expression level correctly.

4. Probe-based Depletion Techniques

In contrast to the unspecific removal of abundant sequences described above, probe-based depletion methods offer the advantage of specifically targeting undesired sequences for removal. This minimizes collateral damage by off-target removal of desired sequences and maintains transcript expression patterns. The downside of this approach is that it requires intricate knowledge of the organism of interest to design the appropriate probe sequences for specific depletion. Probe-based depletion is most commonly used for removal of rRNA transcripts. However, it is also possible to target other abundant sequences for depletion. There are various flavors of probe-based techniques that we will cover in the following section.

Hybridization / Capture Techniques

Hybridization / capture-based depletion methods use a set of affinity probes that specifically target rRNA sequences. The number and positioning of probes varies depending on the number of species targeted, the complexity of the ribosomal RNA sequences in the target groups, and the compatibility for targeting of degraded RNA. For efficient depletion of rRNA from degraded samples, the density of probes on the targeted sequences needs to be higher as breaks in the target region will impair the hybridization of probes at elevated temperatures. The probes contain affinity tags that allow their capture using magnetic beads with corresponding binding sites. Thus, the number of probes contained in a probe mix is closely related to the binding capacity of the beads used for capture. Increasing the number of probe molecules, e.g., by using a very high frequency of probes or targeting a large group of diverse species can be counter-productive for depletion efficiency as it might overwhelm the capacity of specific binding sites provided by the depletion beads. To ensure optimal results, probes for hybridization / capture methods and depletion beads for commercially available solutions are titrated in the optimal ratio.

In the first step, affinity probes are mixed with total RNA and denatured, thereby facilitating access of probes to highly structured target sequences. Hybridization is performed at an elevated temperature to ensure specific binding and to minimize undesired off-target depletion. Depletion beads are used to remove probes that are hybridized to ribosomal RNA from the solution. A final purification step removes all reaction components and recovers pure depleted RNA for downstream applications (Fig. 3).

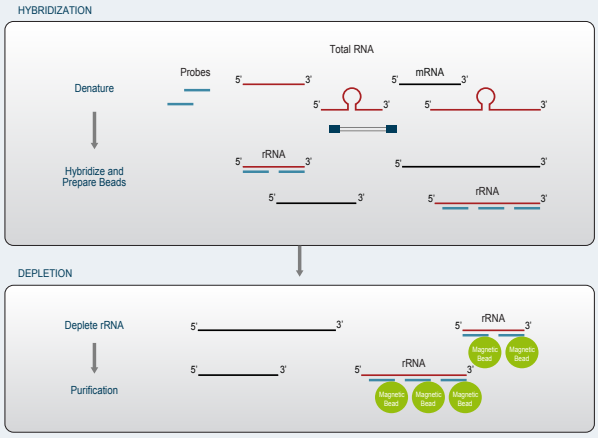


Figure 3 | Depletion of rRNA using specific probes for hybridization / capture. The workflow is adapted from Lexogen's RiboCop Depletion Kits.

Hybridization / capture methods do not rely on enzymatic reactions and therefore these methods leave full-length transcripts intact for downstream processing. They are particularly useful for challenging applications, such as RiboSeq 4, and minimize unspecific RNA degradation.

Here at Lexogen, we aim to provide complete workflow solutions by integrating selection and depletion kits with our innovative library preps. To learn more about our solutions, visit our blog article.

RNase H-mediated Depletion

RNase H is an endoribonuclease that specifically cleaves the RNA molecule within an RNA:DNA duplex while keeping the DNA molecule intact. It is often used in probe-based depletion protocols and falls into the class of specific, enzyme-mediated depletion. This depletion method uses specific DNA probes that hybridize to the target molecule, commonly rRNA and / or globin mRNA. The density of the probes used in the reaction can vary. It is even possible to cover the complete transcript multiple times by using partially overlapping, or so-called "tiling" probes. For rRNA depletion using RNase H, the DNA probes are hybridized to the rRNA at elevated temperatures. After the probes are bound, the reaction is incubated with RNase H which specifically degrades the rRNA.

Depending on the workflow, it is also possible to use a thermostable variant of RNase H (Hybridase™) and incubate the reaction at a temperature of 65 °C or above. This allows increased stringency of the depletion reaction by minimizing nonspecific hybridization at lower temperatures.

Following RNase H treatment, the reaction is incubated with DNase to remove the oligonucleotide probes before the reaction is purified to remove all enzymes and reaction components and elute the now depleted RNA sample (Fig. 4).

RNase H-based degradation is widely used as the reaction components and oligos are cheap and their number can be increased to cover many species. However, the nature of the method being based on degradation poses the risk of unspecific removal of precious transcripts and also makes it unsuitable for certain applications. For example, RNA-Seq workflows that rely on the addition of DNA-adapters prior to depletion should not undergo RNase H / DNase I treatment, as this will degrade the RNA-DNA fusion molecule as well. Recent findings suggest that RNase H-based approaches may not be suitable for challenging applications such as RiboSeq [4].

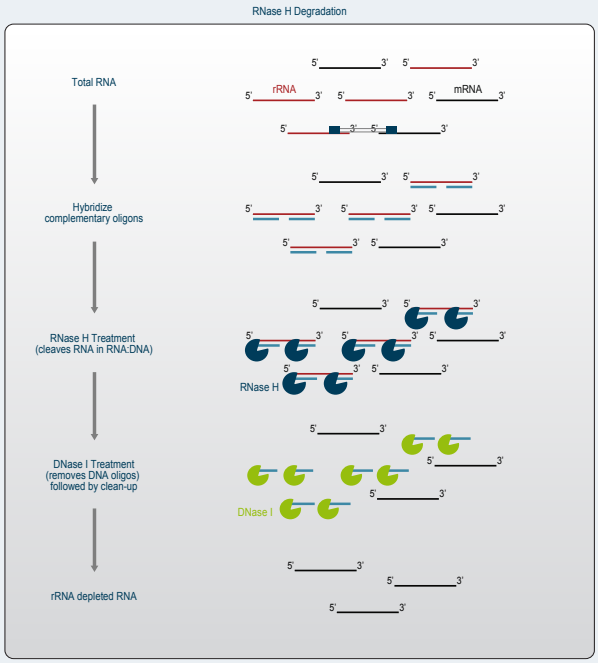


Figure 4 | RNase H-mediated depletion of rRNA using specific probes.

Post-library Prep Depletion by CRISPR-Cas9

CRISPR (clustered regularly interspaced short palindromic repeats) and Cas (CRISPR associated) nucleases, such as Cas9, gained wide popularity in recent years as the "gene scissors". The system has revolutionized genome editing and has a wide range of applications ranging from providing essential molecular biology research tools to solutions for personalized medicine. The groundbreaking discovery by Emmanuelle Charpentier and Jennifer Doudna [5,6] was awarded the Nobel Prize in chemistry in 2020.

The natural CRISPR-Cas system is functional in bacterial adaptive immunity and removes incoming phage DNA without harm to the bacterial genome. The mechanism involves specific cleavage of the foreign DNA by Cas nuclease. Specific guide RNAs which are transcribed from the CRISPR loci bind to the Cas nuclease and guide the enzyme to the complementary target DNA which is then neutralized by Cas-mediated cleavage.

The system can be exploited by providing specialized guide RNAs to target any desired sequence. In RNA-Seq approaches, CRISPR-Cas9 is used together with guide RNAs targeting rRNA sequences or other abundant sequences and can conveniently be used after the library preparation is completed.

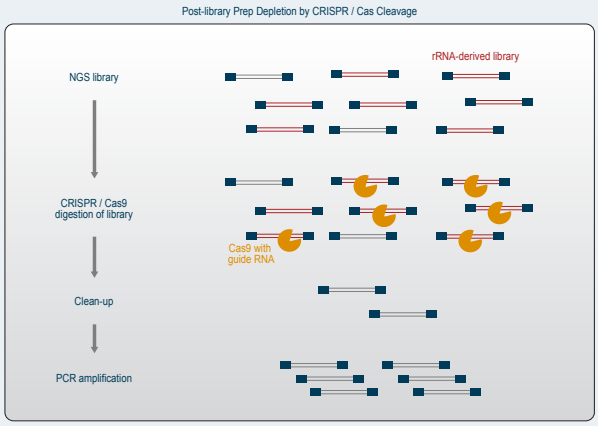


Figure 5 | CRISPR-Cas9 directed post-library prep depletion of rRNA and abundant sequences.

The ready-to-sequence libraries are incubated with Cas9 nuclease that has been pre-complexed with the specific rRNA-guides (Fig. 5). All libraries containing any of the targeted sequences are then cleaved within the pool of molecules. The incubation is followed by a clean-up step to remove the cleaved fragments which are shorter than the remaining non-targeted libraries and are no longer tagged with both sequencing adapters. As this process removes the majority of library fragments, the remaining molecules are re-amplified by an additional round of PCR.

Post-library prep depletion is advantageous when working with samples that do not allow depletion prior to library preparation, e.g., when working with very low input amounts or addressing single cells without oligo(dT) priming. It also allows the researcher to deplete final lane mixes, which saves on depletion cost. On the downside, guide RNA design is quite complex, which may limit the customization potential for non-expert users. Further, libraries for sequencing need to undergo two rounds of PCR which may increase amplification bias. To preserve the guides, storage at -80 °C is advised.

5. In-prep Depletion using Target-specific Blockers

In-prep depletion solutions use sequence-specific blocker oligos that prevent the generation of ready-to-sequence libraries from undesired sequences. Blocker oligos can be used during the reverse transcription step to prevent first strand cDNA generation from these sequences, or they can be added during second strand synthesis to prevent generation of the complementary cDNA strand. To prevent polymerization and strand-displacement activity of the enzymes catalyzing these reactions, blocker oligos are heavily modified to ensure their inertness.

The use of blocker oligos during second strand synthesis allows blocking of abundant sequences also in 3' mRNA-Seq methods. This increases complexity from samples containing abundant mRNAs and elevates the gene detection capacity without the need to increase sequencing depth. One commonly used blocker oligo mix for 3' mRNA-Seq allows removal of reads derived from globin mRNA for Cost-Efficient Gene Expression Analysis in Whole Blood Samples.

6. Exclusion of Abundant Transcripts during sRNA Extraction

Small RNAs (sRNAs) are essential regulators of gene expression and involved in regulatory pathways, such as cancer, inflammation, and development. Isolation with column-based sRNA extraction kits does not confer any specificity since all RNAs below a common threshold of 200 nucleotides are purified. As a result, the majority of sRNA-Seq reads usually correspond to non-functional RNA degradation fragments mostly derived from rRNA and tRNA. To concentrate the sequencing reads on sRNAs, additional selection methods are used, such as size selection by gel extraction or chemical treatment. These methods are extremely laborious, time intensive, and are associated with sample losses. A fast and convenient extraction method, called TraPR, was recently developed by a group of sRNA researchers at ETH Zurich [7]. This method uses a convenient column-based centrifugation step prior to RNA extraction that enriches sRNAs in their functional protein-bound form and removes any free RNA and DNA from the sample. Subsequent RNA extraction then results in a pure fraction of functional sRNA without undesired rRNA fragments, which ultimately saves sequencing space and allows multiplexing of more samples. To find out more about the technology, please see our short RNA Expertise video on TraPR.

Literature:

1. Shagina, I., Bogdanova, E., Mamedov, I.Z., Lebedev, Y., Lukyanov, S., and Shagin, D. (2010) Normalization of genomic DNA using duplex-specific nuclease. *Biotechniques*. 48:455-9. DOI: 10.2144/000113422.

2. Bogdanova, E.A., Shagin, D.A., and Lukyanov, S.A. (2008) Normalization of full-length enriched cDNA. *Mol Biosyst*. 4:205-212. DOI: 10.1039/b715110c.

3. Yi, H., Cho, Y.J., Won, S., Lee, J.E., Jin Yu, H., Kim, S., Schroth, G.P., Luo, S., and Chun, J. (2011) Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. *Nucleic Acids Res.* 39:e140. DOI: 10.1093/nar/gkr617.

4. Zinshteyn, B., Wangen, J. R., Hua, B., and Green, R. (2020) Nuclease-mediated depletion biases in ribosome footprint profiling libraries. *RNA* 26: 1481-1488 DOI: 10.1261/rna.075523.120.

5. Deltcheva, E., Chylinski, K., Sharma, C. et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. 2011 *Nature* 471, 602–607 DOI: 10.1038/nature09886.

6. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E.(2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337: 816-821. DOI: 10.1126/science.1225829.

7. Grentzinger T., Oberlin, S., Schott, G., et. al. (2020) A universal method for the rapid isolation of all known classes of functional small RNAs. *Nucleic Acids Res.*, DOI: 10.1093/nar/gkaa472.

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.

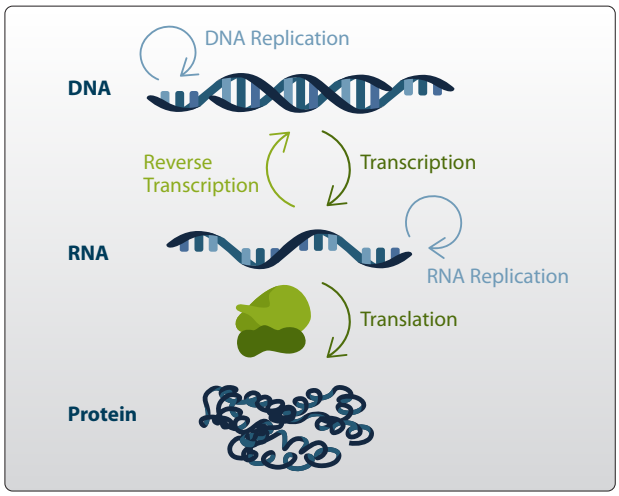


Figure 1 | Central Dogma of Molecular Biology.

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.

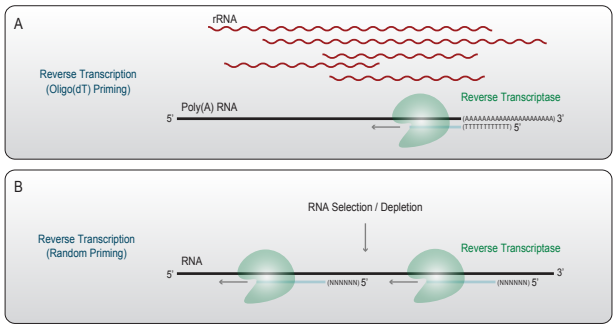


Figure 2 | Reverse Transcription generates cDNA from RNA template molecules. A) Oligo(dT)-primed reverse transcription does not require rRNA depletion or mRNA selection due to the primer annealing specifically to the 3′ poly(A) tail of mRNAs. B) Random-primed reverse transcription uses pre-selected or 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).

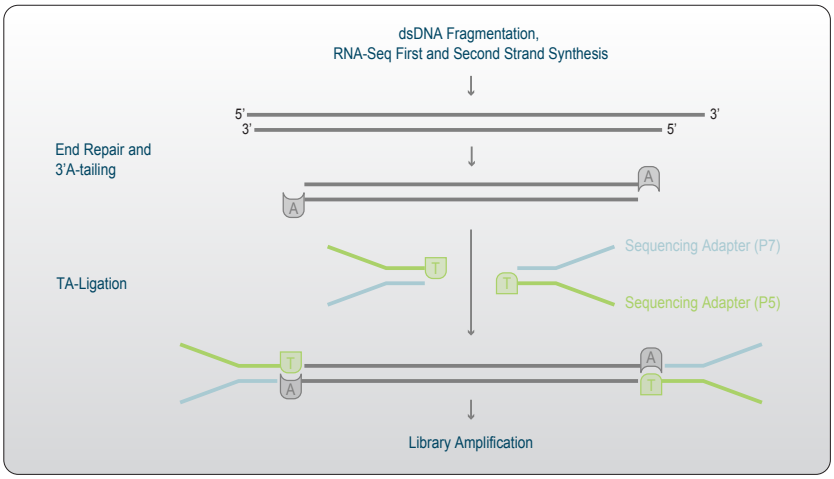


Figure 3 | End Repair, A-tailing and TA ligation. After first and second strand synthesis, single stranded overhangs are removed and the 3′ ends are adenylated. Pre-annealed partially double-stranded sequencing adapters with 3′ T-overhangs are ligated to the A-tailed inserts. Sequencing adapters consist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

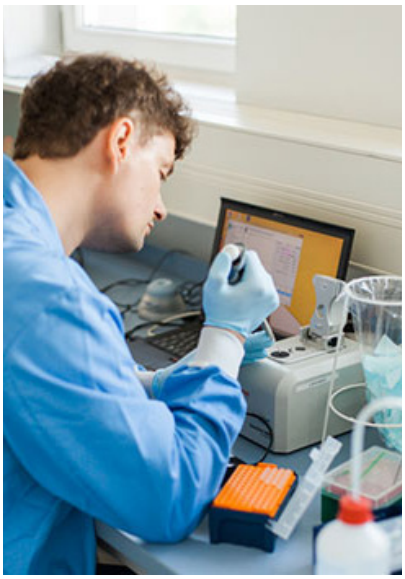
One important aspect for PCR amplification, especially during NGS library preparation, is to avoid entering the plateau stage. Overcycling, i.e., the application of too many PCR cycles, can have a negative impact on NGS data quality, as outlined in Chapter 2 focusing on the sequencing process.

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

What are Unique Molecular Identifiers (UMIs) and Why do We Need Them?

In RNA-Seq experiments, the ultimate goal is to accurately quantify the abundance of RNA transcripts in each sample. During the library generation process, PCR is used to amplify, or copy, transcripts so they can be abundant enough for both quality control and sequencing. During the amplification process, copies are made from identical fragments of the original molecule. As these copies are indistinguishable, it is extremely challenging to determine the original number of molecules in the sample. The use of Unique Molecular Identifiers (UMIs) is an established solution to quantify these original molecules, especially in low-input experiments such as single-cell RNA-Seq.



1. What are UMIs?

UMIs, also known as Molecular Barcodes or Random Barcodes, consist of short random nucleotide sequences which are added to each molecule in a sample as a unique tag. The UMIs are introduced during library generation before the final library fragment is amplified in the PCR step (Fig. 1). This idea was first implemented in an iCLIP protocol[1], a cross-linking and immuno-precipitation method for studying specific protein-RNA interactions.

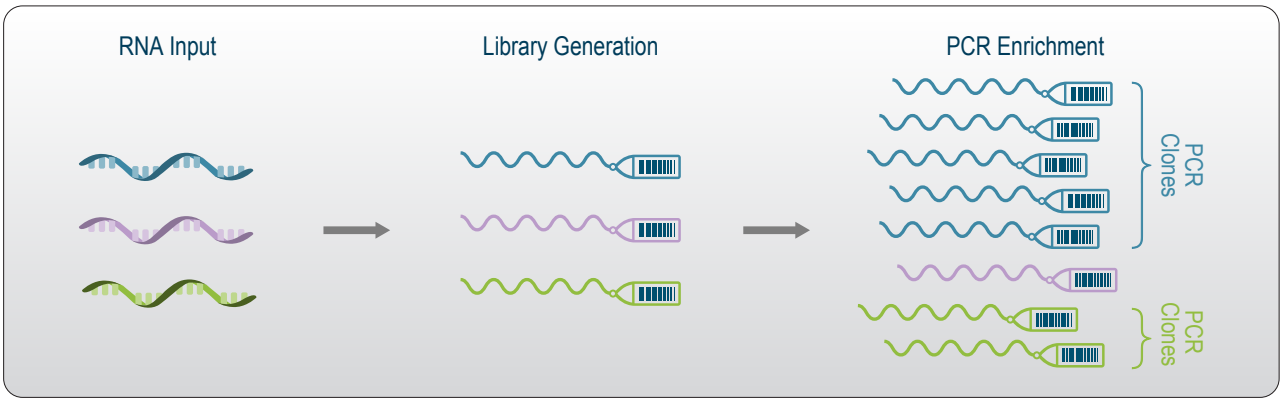


Figure 1 | Individual RNA molecules in each sample are tagged with a unique barcode. These barcodes are copied along with the molecule in the PCR step. Downstream data analysis can then deduplicate the copies, revealing the original ratio of molecules in the sample and eliminating amplification bias.

The primary advantage of including UMIs in a sequencing experiment is to enable the accurate bioinformatic identification of PCR duplicates. Without this capacity, the PCR duplicates can have a detrimental impact on downstream data analysis, especially when amplification biases occurred. Biases in the PCR reaction step can lead to overrepresentation of particular sequences in the final library[2] due to preferential over-amplification. To prevent this bias from further propagation, it has been proposed to remove reads or read pairs with the exact same alignment coordinate, as they are predicted to arise through the PCR amplification of the same molecule[3]. During the later cycles of PCR, error rates increase, and biases can manifest even further (check out the PCR section in LEXICON Chapter 7 for more details).

UMIs therefore ultimately act as tags that allow the accurate identification subsequent removal of PCR duplicates in sequencing data. Thereby, the data quality can be improved by revealing the original number and ratio of molecules in the sample. Despite the ability to rescue some effects of excessive amplification by removing duplicates, biases in the data can be minimized by using the correct PCR cycle number. To learn more about how the optimal cycle number for PCR is determined, visit our blog article.

2. Why are UMIs Useful? Some Applications for UMI RNA-Seq

UMIs enable the quantification of the absolute number of molecules in a sample without the need to detect each individual molecule or identify the number of copies made from them. While measuring the number of copies of each sequence is challenging, counting the number of distinct UMI sequences is easier, and this information does not get lost during the amplification process (Fig. 2). Further, normalization of such RNA-Seq datasets can be performed without the loss of accuracy[4].

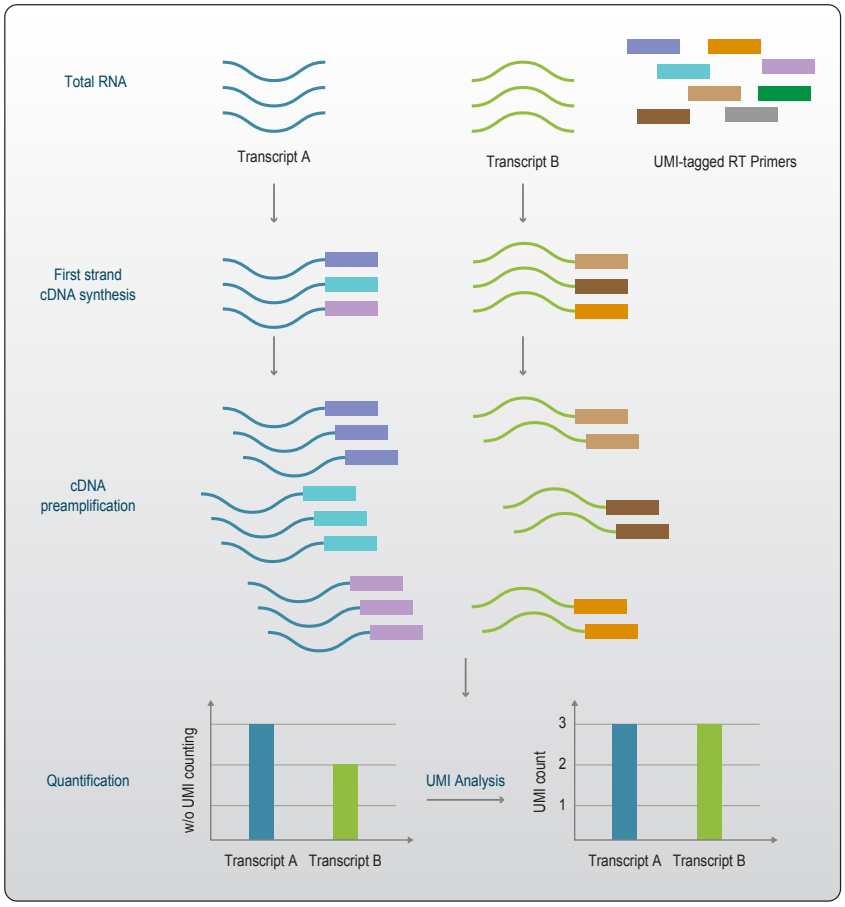


Figure 2 | Transcript level quantification with UMIs. Transcripts or cDNAs are tagged with UMIs in an early step of library generation. The UMI sequences can then be used for quantification of the number of molecules that were originally present in the sample. UMIs can thus control for amplification biases associated with PCR-based sample preparation. Adapted from [5].

UMIs for Transcript / Gene Quantification

Analyzing UMIs is a convenient method of detection and measurement of the abundance of individual molecules present in a complex sample mixture even without an amplification step. Further, different RNA species are present in the sample at different concentrations. It was estimated that differences in concentration between high and low abundant mRNAs in a cell or tissue may vary within 6 to 10 orders of magnitude. These differences in concentrations make the molecular counting procedure via sequencing difficult[4] and can benefit from the use of UMIs. Thus, UMIs may be utilized in any sequencing method, where confident identification of duplicates by alignment coordinate is not possible or where accurate quantification is required. The UMI method could be applied to count all types of molecules or particles such as viruses, proteins, and in methods like ChiP-Seq, karyotyping and others[4].

UMIs for Targeted Sequencing Approaches

Another application benefiting from using UMIs is targeted RNA-Seq, i.e., when libraries are prepared from more restricted regions of the transcript. In these cases, there is a higher chance that identical priming occurs on unique transcripts or first strand cDNA molecules than when using protocols for whole transcriptome RNA-Seq. This results in sequencing reads with identical mapping coordinates and sequences. Without UMIs, these reads may not be quantified correctly, resulting in inaccurate read count data. Including UMIs during library generation, however, clearly distinguishes unique priming events from PCR duplicates and allows for accurate quantification of sequencing reads.

Targeted sequencing can also be used to assess rare variants carrying mutations that can cause a variety of diseases, and are of particular importance in cancer and oncology. Due to the fact, that several steps of the RNA-sequencing workflow can introduce errors, the identification of true rare mutations present in the original RNA molecule can be quite challenging. UMIs are particularly useful to discriminate between errors introduced by the workflow and mutations present in the original molecule[6]. Variants or mutations are considered "true" when they are identical within the individual reads carrying the same UMI and between reads with different UMIs (Fig. 3). Finding the same mutation in reads with different UMIs can further be used to exclude systematic errors introduced by reverse transcription in the first strand synthesis step.

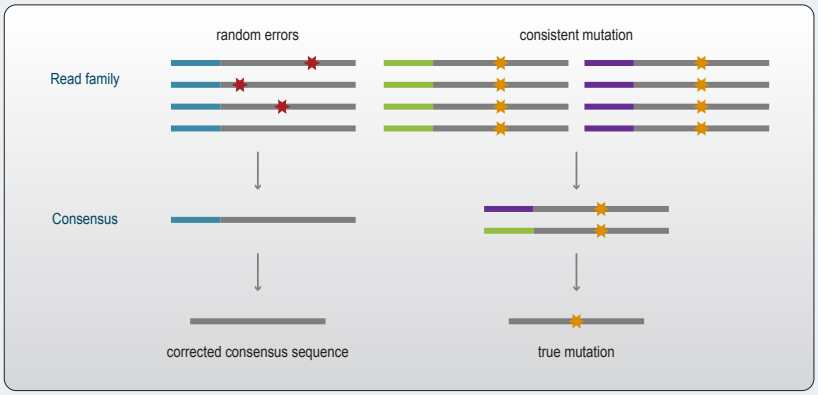


Figure 3 | Alignment of read families sorted by UMIs (color-coded) enables the discrimination of rare variants from random errors introduced during the library preparation and sequencing workflow. True mutations occur throughout the reads carrying the same UMI and can be seen also in reads containing a different UMI. (Adapted from [6]).

3. UMIs in Single-cell Sequencing

UMIs were shown to reduce the amplification noise in single-cell studies[7] and here, UMIs are particularly useful. The amplification bias constitutes one of the main challenges for single-cell experiments and many protocols use multiple, consecutive PCR amplification steps.

A typical single mammalian cell contains approximately 10^5 – 10^6 mRNA molecules and the human cell atlas determined ~11,000 detectable genes in various human cell lines. As genes can be expressed by multiple transcript isoforms differing in their transcriptional start and end sites, exon / intron composition, and expression level, the quantification of transcripts in single cells is particularly challenging. For single-cell experiments, the quantification of genes is often used as it provides an easier and more accurate quantification. In any case, to estimate the number of genes or transcripts expressed in a single cell, UMIs are crucial.

4. How Many Different UMIs are Needed?

UMIs will reflect molecule counts only if the number of available distinct tags is substantially larger than the typical number of identical molecules. The random sequence composition of the UMIs ensures that every library fragment-UMI combination is unique. For this, a large number of random UMI sequences needs to be available.

As an example, for UMI sequences of a length of 10 random nucleotides, as present in QuantSeq-Pool, 4^{10} or 1,048,576 different UMI sequences are used.

It is important to note that the incorporation of the UMI into any library preparation does not interfere with the RNA-seq process, and similar counts of reads mapping to each gene were seen in both UMI-tagged and untagged samples[4]. Since UMIs are agnostic to the library generation chemistry, they are compatible with any indexing strategy, using single or dual indexing and / or additional sample-barcodes (also termed inline indices or sample indices). For more information on indexing, stay tuned for our next Chapter about Indexing Strategies and Solutions.

5. UMIs and the Absolute Truth – Implications for Data Analysis

A fundamental assumption in RNA-Seq has been that library fragments sharing a UMI sequence and read mapping locus were derived from the same initial input molecule. However, it is now clear that a fraction of sequencing reads sharing the same unique molecular identifier would map to different, but closely spaced locations. Due to errors occurring during the sequencing process, the mapping coordinates are not always precise (see Chapter 2 – Next Generation Sequencing for more details).

This imprecision may be misleading for the commonly used analysis tools, which eliminate PCR duplicates and perform counting under the assumption that reads with different mapping coordinates are derived from different starting molecules. This could result in overestimating the expression levels of low abundant transcripts by a large factor and highlights the need for an accurate data analysis pipeline for UMIs in RNA-Seq projects[8].

Some UMI data analysis tools also utilize error correcting functions to account for errors that alter the UMI sequence. For accurate quantification it is beneficial to trace back the erroneous UMI to its parent sequence (Fig. 4).

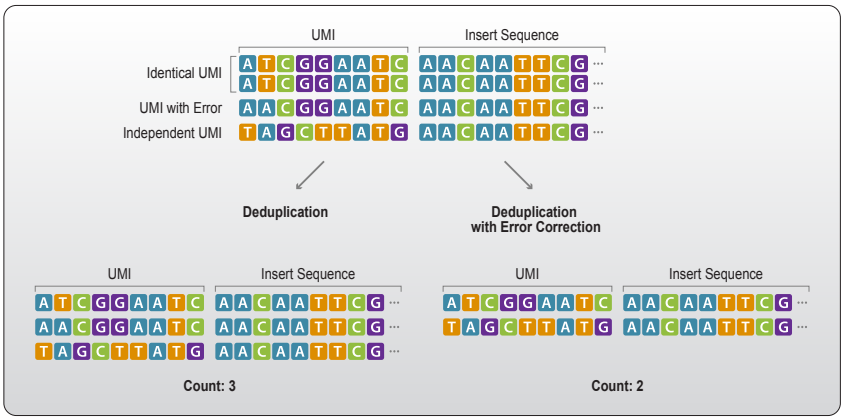


Figure 4 | Errors within the UMI sequence can be corrected for more accurate quantification. Erroneous UMIs can be identified and reverted to the original sequence prior to deduplication. This is done by assessing the distance between the individual sequences and the proposed parental sequence. Deduplication is then based on the parental UMI sequence and also removes reads with slight alterations in the UMI sequence (e.g., caused by nucleotide mis-incorporation).

The knowledge that read mapping coordinates may shift by few bases can also be integrated into the analysis. Stay tuned to find out more about analyzing your RNA-Seq data in one of our upcoming chapters.

6. Some Actionable Advice: When are UMIs Useful for RNA-Seq Library Preps?

UMIs are mostly used to remove PCR duplicates with the aim to reduce amplification bias and to estimate how many genes / transcripts are expressed in single cells. Therefore, UMIs are most useful when the input amounts are limiting (single-cell level to low input amounts of ≤ 10 ng total RNA), while UMIs may not offer a clear benefit for higher input amounts as the number of RNA molecules in this case exceeds the number of possible UMI sequences. UMIs can also indicate over-sequencing, i.e., when the sequencing depth is too high in comparison to the library complexity. While over-sequencing is not harmful *per se*, avoiding over-sequencing reduces costs and frees up sequencing space that can be used to include more replicates to increase the statistical power. UMIs can also help to estimate accessible transcripts, e.g., RNA derived from FFPE samples is very heterogeneous and the number of accessible transcripts varies from sample to sample due to cross-linking.

Library preparation methods that contain built-in UMIs are versatile and can be used for all samples. Processing and deduplicating UMIs is usually optional and can be omitted when working with high input amounts to save computational resources.

UMIs in Lexogen Library Preps

If you would like to use UMIs in your library prep they should ideally be added as early as possible in the process and in any case before the PCR amplification step. The library preparation method of choice thereby defines how and when the UMI is added most efficiently. UMIs can be introduced in the first step during the reverse transcription, e.g., the QuantSeq-Pool 3′ mRNA Library Prep Kit includes UMI sequences as part of the oligo(dT)-primers. It is also possible to add UMIs in the second step, e.g., during second strand synthesis (QuantSeq with UMI module), or during the linker ligation step in CORALL.

UMIs can also be included in an oligo for template-switching if this method is used to generate a second strand. Further, UMIs can be contained in the double-stranded full-length or partial Illumina-adapters which are ligated to double-stranded cDNAs prior to the PCR step.

Literature:

1. König, J., Zarnack, K., Rot, G., et al., (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol.* 17:909-915. Epub. PMID: 20601959; PMCID: PMC3000544. DOI: 10.1038/nsmb.1838

2. Aird, D., Ross, M.G., Chen, WS. et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 12, R18. DOI: 10.1186/gb-2011-12-2-r18

3. Sims, D., Sudbery, I., Ilott, N.E., et al., (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet.* 15:121-132. DOI: 10.1038/nrg3642. PMID: 24434847

4. Kivioja, T., Vähärautio, A., Karlsson, K. et al. (2012). Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* 9, 72–74, DOI: 10.1038/nmeth.1778

5. Kolodziejczyk, A.A., and Lönnberg, T. (2018) Global and targeted approaches to single-cell transcriptome characterization, *Briefings in Functional Genomics*,17: 209- 219, DOI: 10.1093/bfgp/elx025

6. Roloff, G. W., Lai, C., Hourigan, C. S., et al. (2017) Technical advances in the measurement of residual disease in acute myeloid leukemia. *Journal of clinical medicine*, 6: 87. DOI:10.3390/jcm6090087

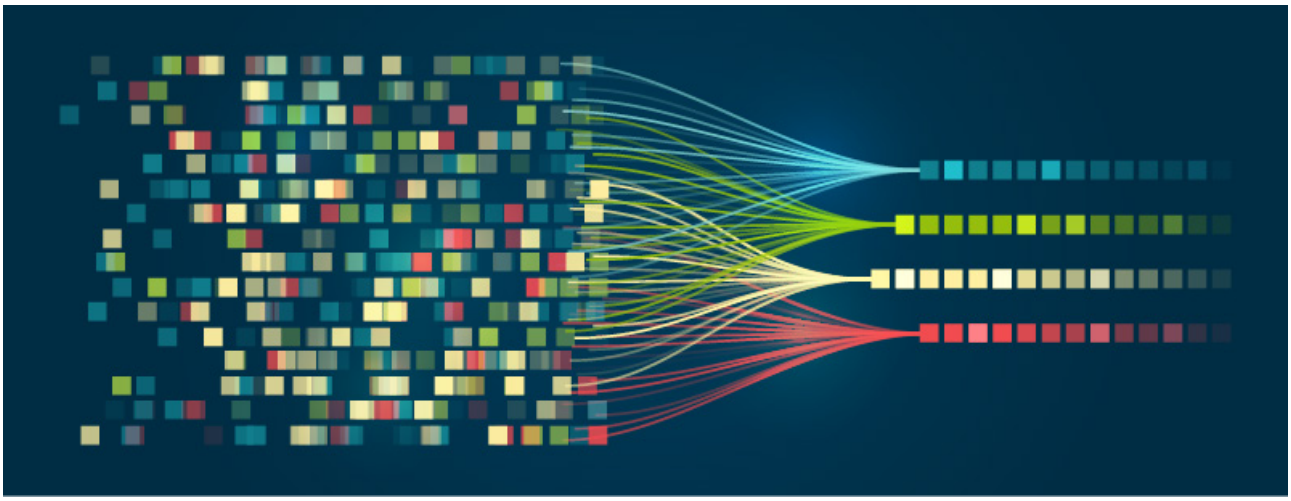
7. Islam, S., Zeisel, A., Joost, S. et al. (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* 11, 163–166, DOI: 10.1038/nmeth.2772

8. Sena, J.A., Galotto, G., Devitt, N.P. et al. (2018) Unique Molecular Identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis. *Sci Rep* 8, 13121, DOI: 10.1038/s41598-018-31064-7

human cell atlas: https://www.proteinatlas.org/humanproteome/cell/cell+line

Indexing Strategies and Solutions

In our previous Chapter 8, we introduced Unique Molecular Identifiers (UMIs) as tags to mark each individual molecule within a sample. In this chapter of the RNA LEXICON, we will focus on a different kind of tag, namely indices. Indices specifically mark each sample in a sequencing experiment and allow simultaneous analysis of many samples in one sequencing run.



1. Sample Multiplexing

High-throughput sequencers produce billions of reads in a single run, heavily outweighing the read depth requirement for single samples which typically lies between 1 M and 100 M reads. Therefore, it is desirable to combine (or "multiplex") libraries from various samples or experiments in one sequencing run. For multiplex sequencing, defined index sequences are added to each library during the Next-Generation Sequencing (NGS) library generation workflow. Each individual molecule generated from an initial RNA sample will have the same index. In contrast, molecules generated from other samples will be tagged with different indices. After sequencing, each read can be identified and associated to the sample it derived from based on the index sequence with which it was tagged. The index tags are typically short defined sequences between 6 – 12 nucleotides. These tags are then read out during the sequencing run.

There are two main strategies for indexing which are commonly used: inline indexing (or sample-barcoding) and multiplex indexing which we will explore in the following sections.

2. Inline Indexing / Sample-barcoding

Inline indices or sample-barcodes are located between the sequencing adapter and the insert (Fig. 1 shows an inline index located at the beginning of Read 2). Due to their positioning, inline indices are part of the insert read, and must be read out either in Sequencing Read 1 or Read 2. Consequently, the read length available for sequencing the actual insert will be reduced by the length of the inline index.

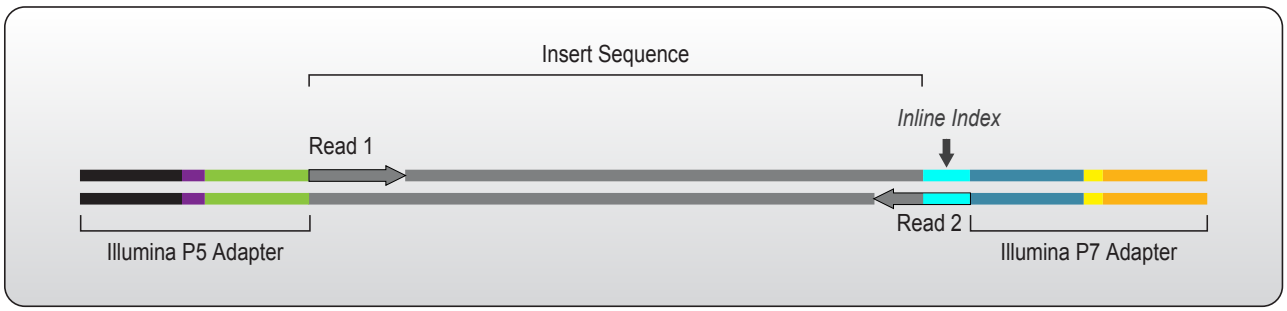


Figure 1 | Inline indices are commonly located at the beginning of Read 2. Read out of inline indices occurs during insert reads and is independent from the multiplex indices located within the Illumina adapter sequences. It is also possible that inline indices are located at the beginning of Read 1 (not shown).

Inline indices or sample-barcodes are commonly introduced in the first step. The index sequence is added to the reverse transcription primer and is therefore mostly located at the beginning of Read 2. Thus, libraries containing inline indices commonly require paired-end sequencing with at least a partial read-out of Read 2.

These indices are commonly used in applications requiring ultra-high throughput. As inline indices are added to the molecules directly in the first step, they allow combination of all indexed samples for subsequent reaction steps. As a result, hundreds of samples can be handled in parallel and multiplexing capacity can be increased to process thousands of samples in one experiment, as exemplified by the QuantSeq-Pool workflow (Fig. 2). Therefore, inline indexing strategies are often pursued for high-throughput screening experiments and for massive single-cell sequencing studies.

Using library preps containing inline indices is not only a convenient way to increase sample throughput, but also saves a lot of consumables as the samples are pooled early and processed in batch. This also effectively shortens hands-on time and can decrease technical variance. For an introduction to sample-barcoded 3' mRNA-Seq check out our RNA EXPERTise video on QuantSeq-Pool.

3. Multiplex Indexing

Continuous improvements in the NGS technology are aimed towards increasing sequencing speed and data output for massive sample throughput. A key to utilizing this increased capacity is multiplex indexing. Just like inline indexing, multiplex indexing allows multiple libraries to be sequenced simultaneously. In contrast to inline indices, multiplex indices are located within the common sequencing adapters and require designated Index Reads to be assessed (Fig. 3). Thus, multiplex indices do not have an impact on the insert read length.

As multiplex indices are part of the common sequencing adapters, they are introduced at a later step in library generation, either during adapter ligation or during the final PCR amplification step.

Multiplex indexing comes in different flavors: single indexing where only Index 1 (the i7 index) is used, and dual indexing that uses both, Index 1 and Index 2 (the i5 index) either in combinatorial mode or as unique index sequence pairs.

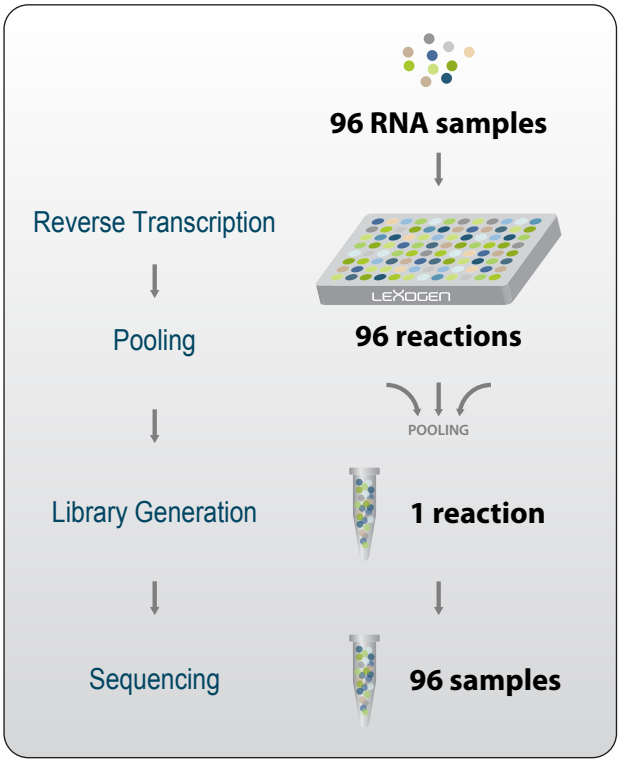


Figure 2 | Inline indexing (sample-barcoding) allows early pooling thereby streamlining the complete workflow. This enables significant savings for consumable and effectively shortens the overall hands-on time to complete library preparation. Additionally, throughput can be upscaled to tens of thousands of samples.

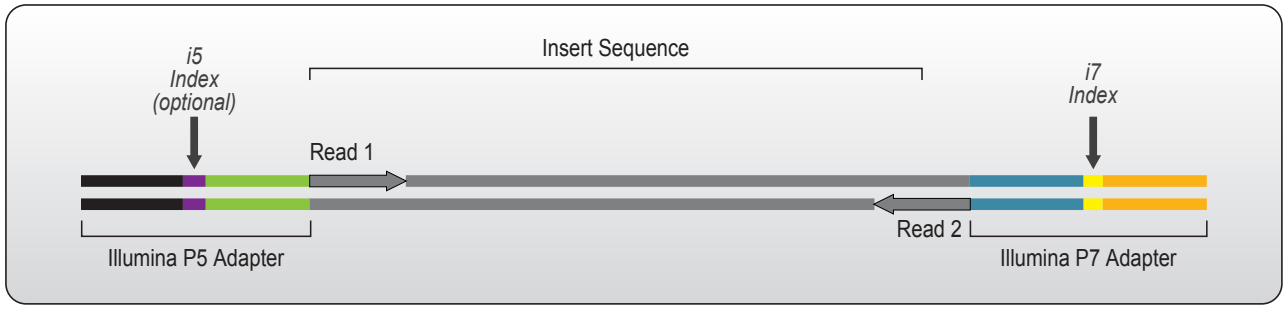


Figure 3 | Multiplex indices are located within the Illumina Adapters. Dedicated Index Sequence Reads are required to assess multiplex indices. Different indexing strategies can be applied: single indexing only uses the i7 index (Index 1), while dual indexing uses both, the i7 (Index 1) and the i5 index (Index 2).

Single Indexing

Sample multiplexing increases sequencing throughput and scalability. However, researchers have since realized that errors occurring in the index sequence also introduce the danger of mis-assignment between the index and the sample that a read originated from. This is especially detrimental for applications that require highly accurate read-out, e.g., when analyzing rare sequence variants – a common application for oncology and cancer research[1].

Single indexing uses the Index 1 sequence as discriminator between different samples sequenced in one run. These sequences are commonly ~8 nucleotides long. While omission of the Index Read 2 shortens the sequencing workflow by ~1 – 2 hours, index sequence errors and the risk of index mis-assignment are the major downsides of single indexing strategies.

Generally, dual indexing strategies are recommended for all NGS experiments. Dual index sequencing requires extra cycles for Index 2 read-out which will prolong the time required for sequencing by ~1 – 2 hours. The reagents for Index 2 read-out are provided in the sequencing cassettes, therefore, researchers do not need to restrict their insert read length to accommodate dual indexing.

Single indexing is still common practice, especially when older sequencers are used, or only a limited number of samples are assessed.

Dual Indexing

Dual indexing has several advantages over single indexing. The largest benefit by far is the increased accuracy for sample assignment and the possibility to correct index sequence errors that would otherwise lead to loss of the read or to mis-assignment to an incorrect sample.

Dual index sequencing offers the chance to identify errors in the index sequence and salvage the reads for later analysis. Once identified, index sequence errors can be corrected when dual indexing is used. The respective second index of the pre-defined pair can thereby be used as a reference point. Without a clear reference point, true error correction is not possible and the chance to falsely correct a given erroneous index sequence is very high.

Dual indexing also allows to multiplex more samples per sequencing run as the number of possible index combinations is tremendously increased, e.g., with 96 different i7 and 96 different i5 indices, a total of 9,216 (96 x 96) index combinations is possible.

Newer instruments and sequencing chemistries have been optimized for ultra-high throughput sequencing to ensure increased data output, faster run times, and cost reduction per run. As a trade-off, more index sequence errors and a higher level of index misassignments were observed when using these new instruments[2]. The use of dual indices, especially in a unique i5 / i7 combination allows to remove any read whose source cannot be unambiguously identified. Thereby, detrimental index mis-assignment can be averted also in highly multiplexed experiments.

4. Multiplex Dual Indexing – Practical Implications

Dual indices can be applied in two different ways – either in a combinatorial or in a non-redundant (= unique) manner. Combinatorial dual indexing uses each individual i5 and i7 index multiple times whereby each combination of these indices is only used once in the experiment (Fig. 4).

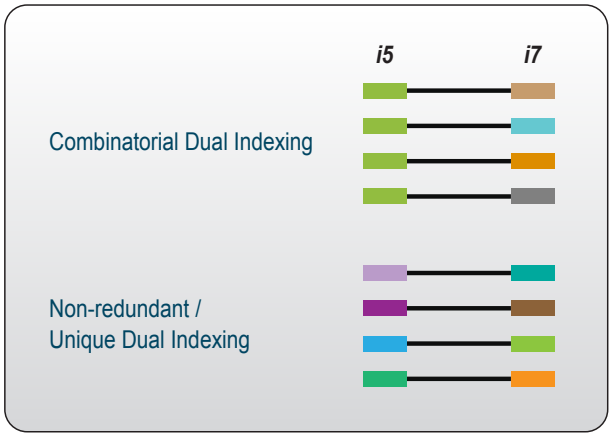


Figure 4 | When using combinatorial dual indexing, every i5 and i7 is used multiple times; therefore, the combinations are unique, the individual indices are not. In contrast, when using unique dual indexing, each i5 and i7 index is used only once; every combination and every index is therefore unique.

This allows a tremendous increase in multiplexing capacity and concomitantly reduces the overall per-sample cost. However, as the barcodes are shared between multiple samples, it is not always possible to unambiguously identify the corresponding sample in case of index sequence errors.

When following a unique dual indexing strategy on the other hand, each individual i5 and i7 index is used only once in the experiment (Fig. 4). As a result, index crosstalk can be dramatically reduced, and index mis-assignment can be prevented[3].

In case of errors, the second index of the pair can be used as a reference point to pinpoint the identity of the original index pair. This ultimately has the potential to salvage a large fraction of otherwise unassigned reads by reverting the erroneous sequence back to the original sequence in a process termed "index error correction". In typical sequencing experiments, ~10 % of the reads cannot be assigned and would therefore be discarded when error correction is not applied.

Unique Dual Indices (UDIs) are recommended for best practice and for the highest possible accuracy for demultiplexing / index assignment.

Advantages of Unique Dual Indexing

- ✓ UDIs increase the accuracy of sample identification by using two unique identifiers.
- ✓ UDIs enable identification of index errors and index-sample-swaps or index hopping. *This is not possible when single indexing is used. As the i5 index is lacking, there is no second reference point to assess which sample the read originated from when the i7 index is changed to an ambiguous sequence, i.e. it could have originated from another sample.*
- ✓ Well-designed UDIs are the basis for index error correction. Index error correction can rescue unassigned reads that would otherwise be discarded. *As a practical example, the ~10 % of discarded reads from a NovaSeq (S4 FlowCell) can account for up to two full NextSeq500 runs, or up to 800 M reads, which can be saved when using UDIs and error correction.*
- ✓ Ultimately, UDIs reduce per-sample costs and maximize sequencing output.

5. Index Sequence Design – From Distances to Indices

Index sequence design is extremely important for the improvement in accuracy that unique dual indexing can offer, and it determines the error correction capacity of the index set. In this section, we will dive into design features and explain what makes an index set truly advanced. One obvious requirement for index sequence design is to provide the necessary color- and nucleotide-balance to ensure a high enough complexity for a smooth sequencing process and signal detection. A major factor that determines the quality of any given index set is the inter-index distance (also referred to as inter-barcode distance).

The inter-index distance is a measure of *dissimilarity* between sequences in a given set. The distance is defined as the number of edit events that are required to transform any one sequence into any other sequence of the same set. The higher the inter-index distance, the more edit events are needed for this transformation. Or in other words: the larger the inter-index or edit distance, the more unlikely it is to create false-positive barcode matches, and the easier it becomes to detect and correct erroneous index sequences.

"Edit events" summarizes all types of errors that can modify the nucleotide composition of a sequence, i.e., nucleotide substitutions, where one base is exchanged for another at the same position or any changes that alter the positioning of nucleotides in the sequence context, such as insertions, where a base is added and deletions where a base is removed (Fig. 5).

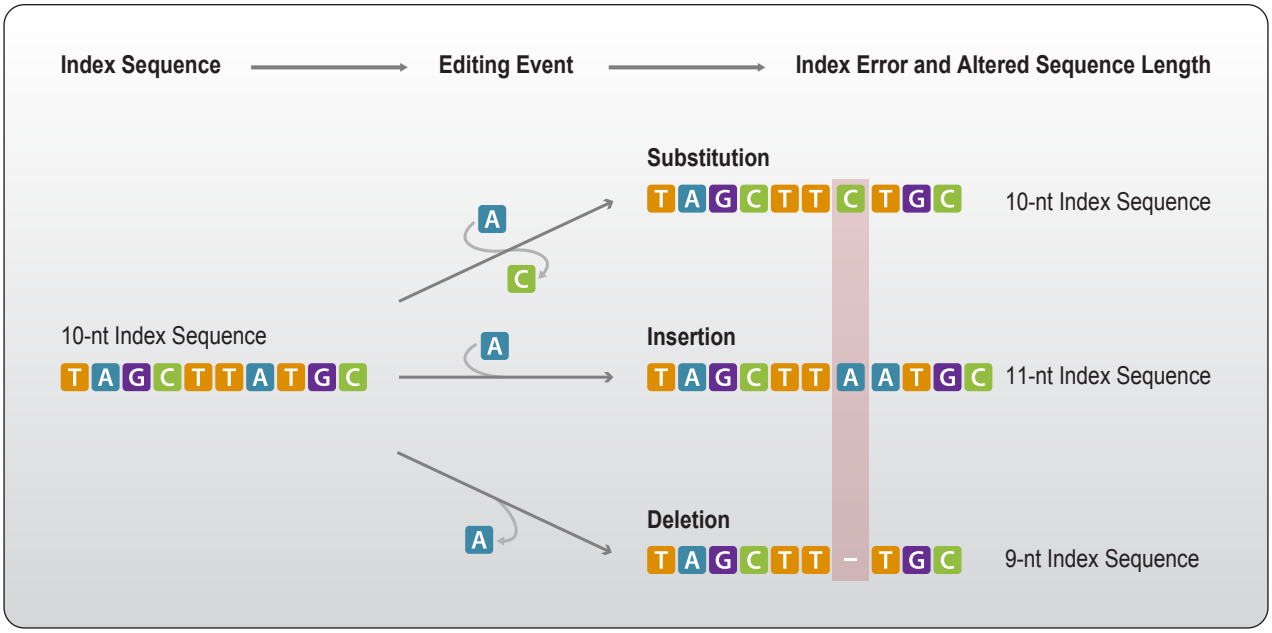


Figure 5 | Edit events change defined into unknown sequences. ***Substitution:*** *a base in the sequence is replaced by another base, e.g., an adenosine is substituted by a cytosine.* ***Insertion:*** *a base, e.g., adenosine, is added at any given position; all following bases are shifted by one position generating a longer sequence.* ***Deletion:*** *a base, e.g., adenosine, is removed at any given position; all following bases are shifted by one position generating a shorter sequence.*

Key Principles for Illumina-compatible Index Sequence Design

- ✓ Color- and nucleotide balance should be considered in the design to ensure efficient sequencing and signal detection on the machine. *Illumina machines using 2-color chemistries (i.e., only two fluorophores are used to distinguish the four different nucleotides) may require a higher nucleotide diversity than machines applying 4-color chemistry and a different fluorophore for each nucleotide.*
- ✓ Index Sequence length: longer index sequences have a higher inter-index distance than shorter index sequences. *Longer sequences possess a more complex sequence space, i.e., more possible nucleotide combinations. Due to the higher number of overall possible sequences the ones with the largest index-distances can be chosen.*
- ✓ The inter-index distance chosen for the design has implications on the types of errors that can be detected and corrected. *The following index distances are commonly used: Hamming distance, Levenshtein distance, and Sequence-Levenshtein distance, as well as modifications thereof. To learn more about these distance types and their implications for index design, see below.*

Hamming Distance

The Hamming distance was introduced by Richard W. Hamming in the 1950s and is a measure for dissimilarity between two strings of characters that are equal in length. In terms of index sequences the Hamming distance can be used to describe the number of positions in which the bases of two index sequences differ. It measures the minimum number of substitutions required to change one sequence into the other.

While the Hamming distance is used for binary strings, it can also be explained as using a codeword scheme. For example, it can be applied to words of equal length. To transform the name "Addison" into "Allison", only two letters need to be exchanged. Therefore, the Hamming distance between Addison and Allison is 2.

As the Hamming distance requires both sequences to be of equal length, deletions and insertions cannot be assessed (see Fig. 5). One downside of not being able to take insertions and deletions into account is that a large Hamming distance does necessarily reflect a large edit distance (Fig. 6).

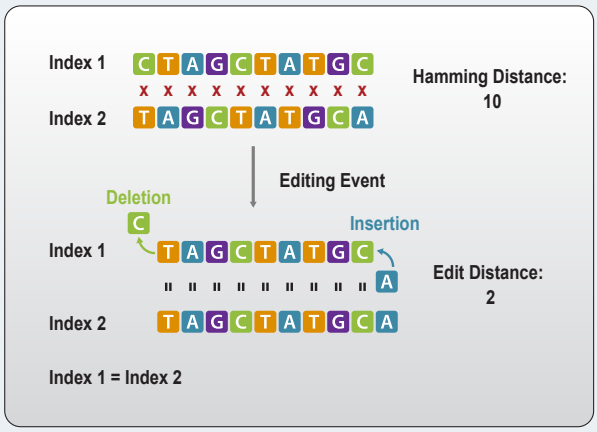


Figure 6 | Insertions and deletions reduce the usability of Hamming-based Index sequences. Two indices that are different from one another by 10 substitutions (Hamming distance =10) can have an edit distance of two, i.e., that a total of two insertions or deletions can turn Index 1 into Index 2 leading to sample misassignment. Adapted from [4].

Therefore, the Hamming distance may not be the most appropriate design parameter to ensure sophisticated index sequence design. Rather, modifications of the Levenshtein distance are used to account for editing events other than substitutions.

Levenshtein Distance

The Levenshtein distance is another string metric to describe the difference between two sequences. It was introduced by Vladimir Levenshtein in the 1960s. The Levenshtein distance between two sequences is defined as the minimum number of single-character edits required to change one sequence into the other. In contrast to the Hamming distance, the Levenshtein distance can also assess substitutions and deletions and allows to compare sequences of variable lengths (Fig. 7).

The Levenshtein distance is more flexible than the Hamming distance and can cover all editing events that can occur at index positions during a sequence workflow. The NGS-specific problem that arises for the classic Levenshtein distance is that during a sequencing experiment, the read-out is fixed. For example, the index read will always be 10 imaging cycles, i.e., 10 nucleotides will be read out even when the index length is changed to 9 nucleotides by deletion or 11 nucleotides by insertion. This means that "non-index nucleotides" will be moved into the sequencing frame in the course of deletions and "index-nucleotides" will be moved out of the sequencing frame upon nucleotide insertions.

As a consequence, the Levenshtein distance as originally conceived is also not optimal for index sequence design.

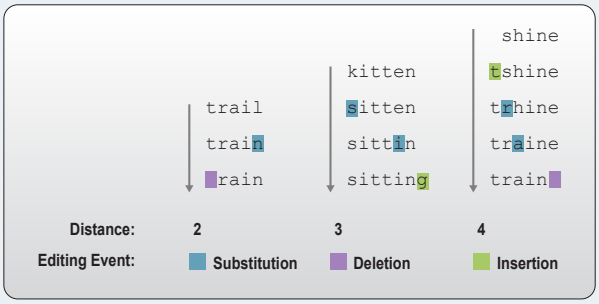


Figure 7 | Levenshtein distance as exemplified by the transition from one word into another. The transition from "trail" to "rain" requires at least 2 edits and thus has a distance of 2. The Levenshtein distance between "kitten" and "sitting" is 3, and between "shine" and "train" it is 4 as at least 4 editing events are required to change one into the other. Editing events are defined as either insertion, deletion, or replacement of a character (substitution).

Here at Lexogen, we strive to improve every step of the sequencing process. Therefore, Lexogen has designed and produced the most sophisticated Unique Dual Index (UDI) Set on the market to date. The result is a versatile, scalable, and nested UDI Set with maximized inter-index distance for all sample sizes. Ultimately, these UDIs enable superior error-correction and allow for tremendous cost savings through maximized sequencing output by rescuing the majority of unassigned reads.

Sequence-Levenshtein Distance

While the length of the index is known per design, the length of the actual observed index in a sequencing experiment can be altered and is thus an unknown variable.

In case of nucleotide deletions, the nucleotides located downstream of the sequencing frame are moved into the index space. If the index length is increased by nucleotide insertions, bases belonging to the original index sequence are moved out of the index space and now precede the nucleotides of the adapter or insert (Fig. 8). These bases will not be seen in the data as the index reads are usually not increased beyond the original index boundaries.

The Sequence-Levenshtein distance is a variation of the original Levenshtein distance described above, it is adapted to account for the sequence context in a continuous flow. It can account for changes caused by appended non-index nucleotides and the resulting shorter distance between the read-out index sequence. Thereby the actual length of the erroneous index can be correctly identified as well as appended nucleotides[5]. Nucleotides that move in and out of the index space can generate sequences with shorter distances to other indices in the set as compared to the original index sequence they are derived from. A large inter-barcode distance based on the Sequence-Levenshtein distance, therefore, does not necessarily guarantee accurate error correction.

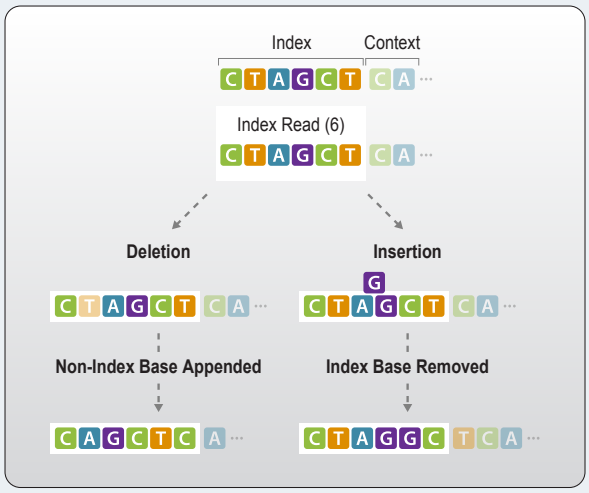


Figure 8 | Impact of deletions and insertions on the index read-out. Upon deletions within the index sequence, non-index bases succeeding the index nucleotides enter the frame of the index read. Insertions within the index sequence leads to index-associated bases to move out of the index read frame. They now precede the downstream sequence, i.e., either the adapter sequence when multiplex indices are used or the insert sequence when inline indices are used.

Further advances in this field improve index sequence design by accounting for the probability of deletions, substitutions, and insertions in sequencing experiments and focusing on the shifts that can be caused at the 3' end of the index sequence[6].

6. The Best of Both Worlds – Combining Indexing Strategies

Combining inline and multiplex indexing allows to take sample multiplexing even further: experiments can be easily scaled up to tens of thousands of samples for ultra-high throughput applications, such as massive screening projects. The combination of inline indices and UDIs in a triple index system enables highly confident sample assignment for more than 36,000 individual samples, e.g., 96 sample barcodes combined with 384 UDIs, 96 x 384 = 36,864 (Fig. 9).

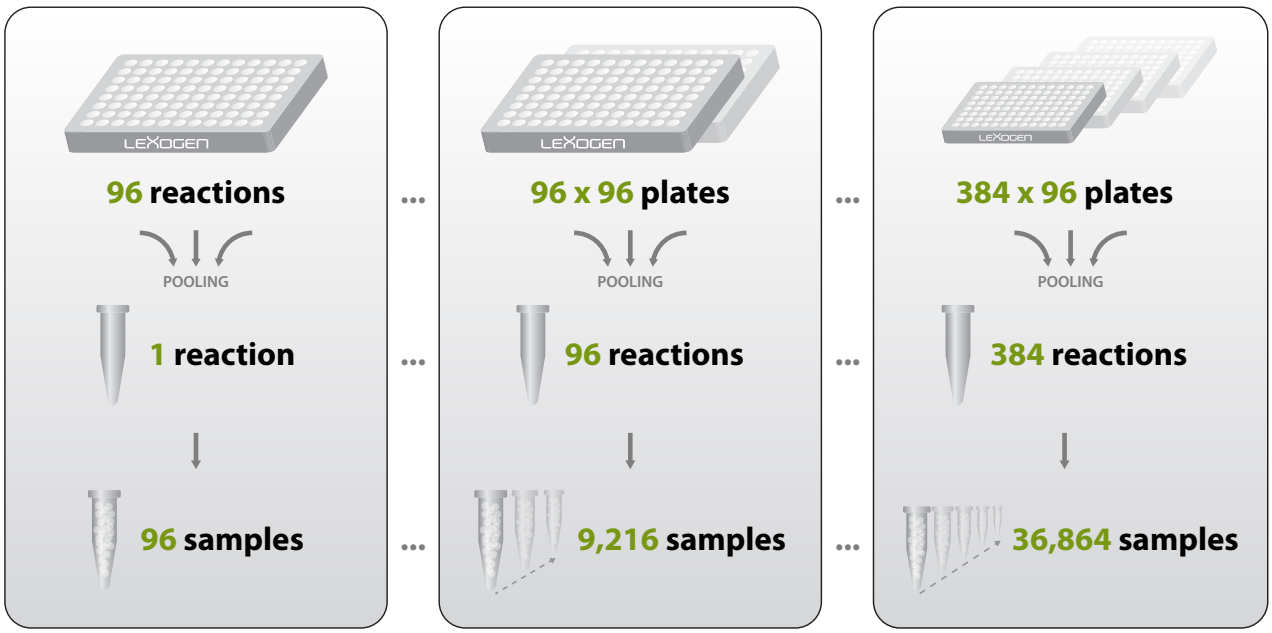


Figure 9 | Highly scalable throughput and confident sample assignment by combining inline indexing (sample-barcoding) with unique dual multiplex indices. For example, combining 96 inline indices with 96 or 384 UDIs allows multiplexing of 9,216 or 36,864 samples and thus provides a cost-efficient solution for large scale screening approaches.

6. Some Actionable Advice – The Connection between Flow Cell Chemistries, Instruments, and Indexing

While Unique Dual Indexing is the gold standard for RNA-Seq, the characteristics of sequencing instruments themselves can impact the run performance in a way that can influence the choice of indexing strategy. By far, the most common sequencing instruments used are those made by Illumina, whose sequencer portfolio utilizes two separate types of flow cells: patterned, and non-patterned (also known as random flow cells).

Non-patterned flow cells have a uniform surface on which cluster generation occurs randomly across the flow cell. Provided the user loads the flow cell with the appropriate concentration, cluster generation will generally succeed without issue (Fig. 10, left). Now, advancements have resulted in the adoption of a patterned flow cell, where the surface of the flow cell is occupied by billions of nano-wells in which the cluster generation is occurring within a known, defined physical space (Fig. 10, right). There are a number of advantages to the patterned flow cell. The cluster generation is more tolerant of a wide range of loading concentrations, since the nano-wells reduce the chance of over-loading the flow cell. Also, because the nano-well locations are known, there is no need to map the cluster sites, saving time during sequencing.

That being said, there is one major drawback to the patterned flow cell, which is the increase in index hopping events[2]. The patterned flow cell uses a different type of sequencing chemistry, dubbed Exclusion Amplification, or ExAmp chemistry. This replaces the bridge amplification method previously used for cluster generation on non-patterned flow cells. In ExAmp chemistry, all of the reagents needed for cluster generation are mixed before the cluster generation occurs, which is the likely cause of index swapping as there are free index primers in the mix ahead of cluster generation. Whereas in the traditional bridge amplification chemistry, these free index primers are removed in a washing step after hybridization of DNA to the flow cell. It is estimated that up to 6 % of reads on patterned flow cells can be affected by index switching, compared to less than 1 % on non-patterned flow cells[7].

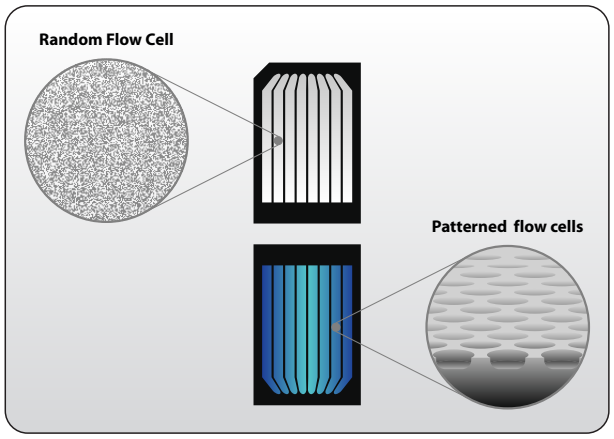


Figure 10 | Random and Patterned flow cell for Illumina Sequencers. Left: Clustering on random / non-patterned flow cells occurs randomly by library molecules binding to flow cell oligos attached to the surface. Clustering is influenced mainly be the loading concentration of libraries. Right: Patterned flow cells are characterized by regularly spaced nano-wells that contain the flow cell oligos. Cluster generation occurs only within the nano-wells making the flow cell less sensitive to overclustering while increasing cluster density and concomitantly data output. Find more information on www.illumina.com.

The benefits of the patterned flow cell are significant, and this is shown in the sweeping adoption of them across the Illumina sequencer family. The ease of loading and shortened sequencing time are major benefits as the amount of multiplexing increases continually. Therefore, it is imperative to use the available tools to mitigate the unavoidable increase in index hopping events introduced by the patterened flow cell technology. For best practice, it is strongly recommended to use the unique dual indexing strategies outlined in this chapter, particularly those with the capacity for error correction, which can rescue a truly substantial quantity of reads.

Literature:

1. Kircher, M., Sawyer, S., & Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic acids research,* 40:e3, DOI: 10.1093/nar/gkr771

2. Illumina White paper. Effects of Index Misassignment on Multiplexing and Downstream Analysis.

3. MacConaill, L.E., Burns, R.T., Nag, A. et al. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics* 19, 30 (2018), DOI: 10.1186/s12864-017-4428-5

4. Faircloth, B.C., and Glenn, T.C. (2012) Not All Sequence Tags Are Created Equal: Designing and Validating Sequence Identification Tags Robust to Indels. *PLoS ONE* 7(8): e42543, DOI: 10.1371/journal.pone.0042543

5. Buschmann, T., and Bystrykh, L.V. (2013) Levenshtein error-correcting barcodes for multiplexed DNA sequencing. *BMC Bioinformatics* 14, 272, DOI: 10.1186/1471-2105-14-272

6. Hawkins, J. A., Jones, S.K., Finkelstein, I.J., and Press, W. H. (2018) Indel-correcting DNA barcodes for high-throughput sequencing. *Proceedings of the National Academy of Sciences*, 115:E6217-E6226, DOI: 10.1073/pnas.1802640115

7. Costello, M., Fleharty, M., Abreu, J. et al. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* 19, 332 (2018), DOI: 10.1186/s12864-018-4703-0

Library Preparation Quality Control and Quantification

In this chapter of RNA LEXICON, we are taking a closer look at quality control methods for library preparations. Quality control measures ensure that your library preparations are as good as can be before you venture to sequencing. As described in our previous chapters, commonly used library preparation methods involve PCR amplification steps to complete the adapter sequences, introduce indices and amplify the final library to a level that allows quantification and quality control using dedicated devices. Quality control aims to accurately determine the library profile, size distribution, and concentration for loading the sequencer. Most experiments aim for an equal read distribution between all samples to ensure comparability of the samples in later analysis. Even though many data analysis tools use normalization prior to the comparison of sample groups, larger differences in read-depth between the groups can cause various unexpected effects. To avoid these complications, equal read distribution is the gold-standard for most RNA-Seq experiments and accurate quantification of the libraries a prerequisite.

1. Quality Control Methods

The analysis of a small volume of the amplified library with micro-capillary electrophoresis has become standard practice for NGS laboratories. Electrophoresis / microfluidics platforms are available from various manufacturers, e.g., Bioanalyzer, Fragment Analyzer, LabChip GX II, or TapeStation. The traces generated on these machines deliver information about the library quantity, size distribution, shape, and the presence of undesired by-products or residual primers (Fig. 1).

In case substantial by-products are visible (e.g., by-products accounting for > 3 % of the final library preparation), it is best to remove them by re-purifying the final lane mix prior to sequencing. As shorter fragments are preferentially amplified, the by-products can otherwise take up a significant amount of sequencing space and reduce the number of useful sequencing reads obtained from an NGS experiment.

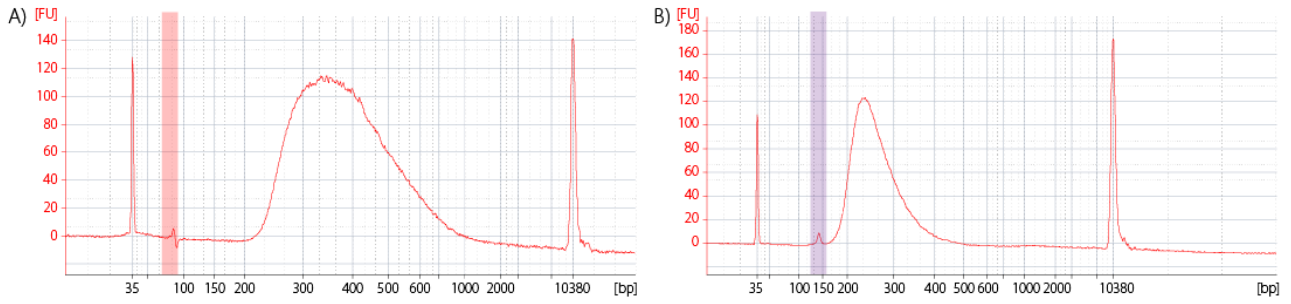


Figure 1 | Bioanalyzer traces of final RNA-Seq Libraries. A) Library with residual primers after the final purification step, highlighted in red. B) Library with adapter dimer (alternative: linker-linker) by-product after the final purification step, highlighted in purple.

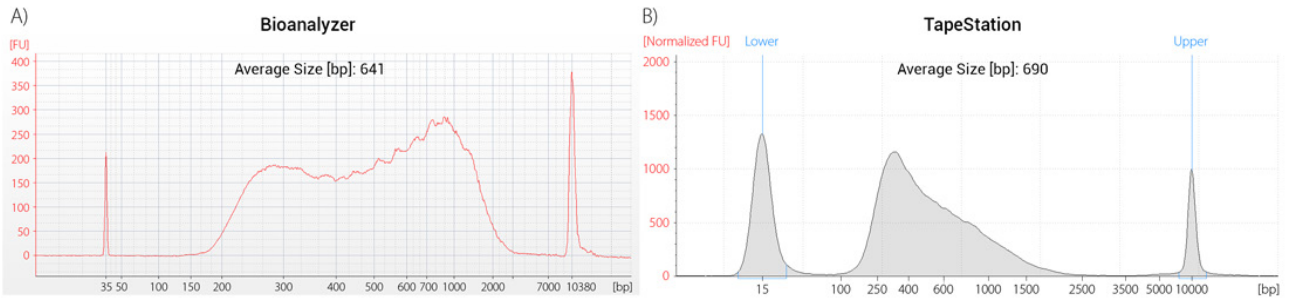


Figure 2 | Machine-specific appearance of library traces. Aliquots of the same library were run on a Bioanalyzer High Sensitivity DNA Chip (A) and on a TapeStation High Sensitivity D5000 ScreenTape (B).

Which of these machines is used for library quality control often depends on the needs and sample throughput of the individual laboratory, e.g., Bioanalyzer is a chip-based system with the capacity to analyze 11-12 samples per run while Fragment Analyzer is plate-based and can handle multiple 96-well plates per run making it the go-to solution for high-throughput NGS laboratories with large scale projects. In addition, the machines differ in their resolution, sensitivity, and dynamic range. While all of them can be used for library QC, input requirements and the appearance of the library trace will vary between instruments (Fig. 2).

While microfluidic devices offer thorough information on the relative size distribution and presence or absence of side products, these methods should be combined with a sensitive quantification assay for more accurate results. For example, library quantification can be performed using benchtop fluorometers with an assay for highly sensitive DNA quantification, e.g., the Qubit dsDNA HS assay is often used.

2. qPCR for Accurate Quantification of Amplifiable Fragments

More accurate library quantification can be achieved with qPCR assays. With these assays, the relative or absolute abundance of amplifiable fragments contained in a ready-to-sequence library is assessed. The qPCR assays use specific primers targeting the adapter regions only present in fully functional library molecules. Thereby, only the fraction of library molecules that are correctly assembled and amplifiable is assayed. The concentration of these fragments is then calculated by comparing Cq or Ct values to a set of known standards (Fig. 3).

While delivering a more accurate quantification, these assays do not supply the user with information regarding library size distribution. Unwanted side-products such as linker-linker artifacts are not discernible from the actual library in the qPCR assay as both will be amplified.

Also, qPCR assays rely on intercalating dyes for quantification, such as SYBR Green I or EvaGreen. These dyes interact non-specifically with double-stranded nucleic acids. The signal strength is proportional to the length of the double-stranded molecule, i.e., the longer a double-stranded fragment, the more dye molecules can bind and interact with it and the stronger the signal will be for this specific library fragment. As a consequence, accurate quantification also requires normalization of the estimated molarities according to the average library lengths. It is therefore highly recommended to combine such an assay for quantification with microcapillary electrophoresis analysis for library size distribution and assessment of by-products that would influence the measurement.

Apart from using qPCR to quantify the final library, the qPCR assay is also a useful tool to quality control the workflow during library preparation, especially when using low input RNA or single-cell library preps where the input cannot undergo quality control as described in Chapter 4.

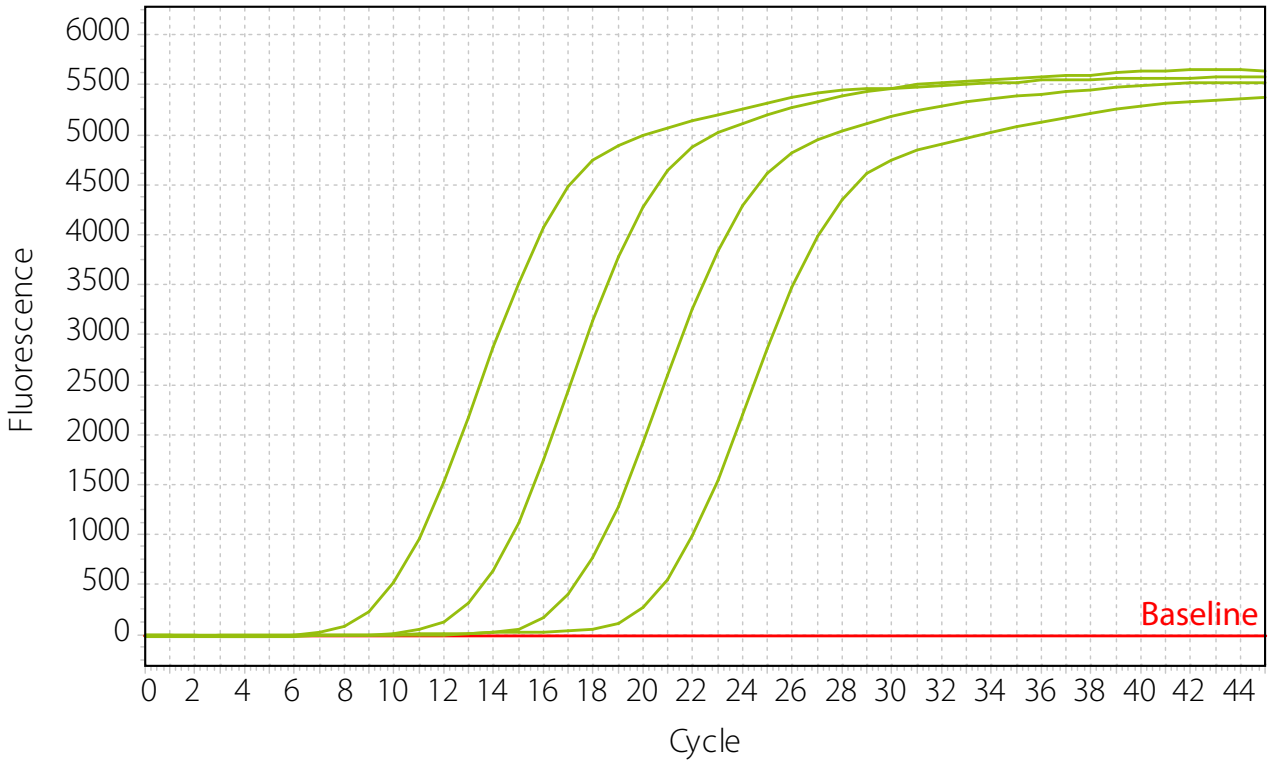


Figure 3 | Library quantification assay using qPCR based on known standards. Standards with defined concentration are assayed in parallel to the libraries to determine the concentration of the unknown libraries. Standard curves are shown (from left to right: high concentrated to low concentrated standard).

What is a Cq or Ct Value?

The Cq value or "quantification cycle value" is defined as **the number of cycles required for the fluorescent signal to exceed the background fluorescence during a qPCR reaction**. This value is also referred to as Ct ("threshold cycle"), Cp ("crossing point") or TOP ("take-off point").

Even though fluorescent dyes used in qPCR are specific to double-stranded products, a considerable amount of background fluorescence is commonly detected. This is also the case when sequence-specific probes are used, however, the background fluorescence might be at a lower level. It is therefore critical to surpass the basal level of the fluorescent signal in order to quantify the amplification product correctly.

The threshold is defined as the fluorescence level above background at which a signal can be detected. **The C_q value is the PCR cycle number** at which your sample's fluorescence curve reaches the threshold (Fig. 4). It therefore reflects the level of amplification that was required to detect the sample and as such it is inversely correlated with the amount of template inserted into the reaction.

Samples with low Cq values reach the threshold fast, i.e., they require a lower level of amplification to surpass the background as the concentration of the PCR template in the reaction was high to begin with. Samples with high Cq values required more rounds of exponential amplification and therefore contained less of the target molecules.

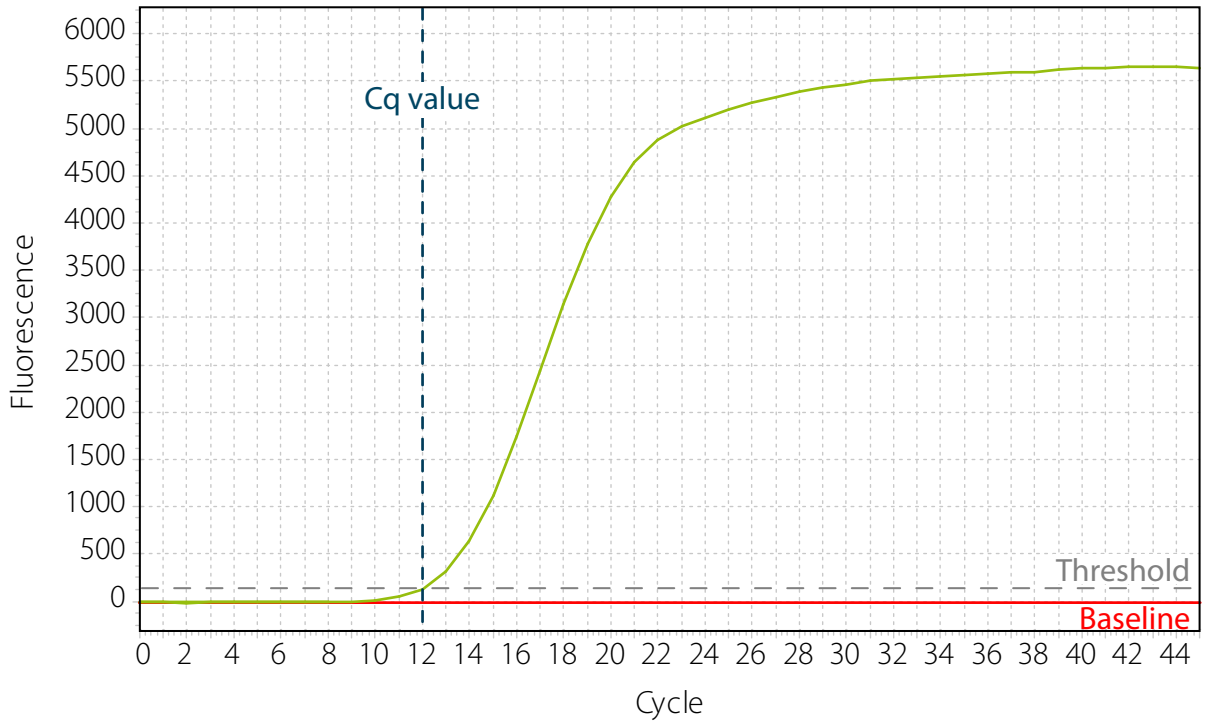


Figure 4 | Cq value and threshold of a qPCR amplification curve.

3. qPCR for Quality Control during Library Generation

Having covered post-library preparation quality control mechanisms, we will briefly focus on quality control steps that can be used during library generation to make sure libraries of the highest possible quality are obtained. In Chapter 4 and Chapter 5 we have discussed how microfluidics assays are used to assess the quality of the RNA after extraction and how genomic DNA can be detected in your sample.

Apart from RNA quality control, qPCR is used as a tool to control any pre-processing steps, such as mRNA-selection and rRNA depletion and the subsequent library generation steps. It helps to assess the efficiency and consistency of the individual processing steps, i.e., technical replicates should show the same Cq / Ct values and behave identically. Large variations between technical replicates thus point to inconsistencies in handling, environment, or the protocol itself.

For long-term and large-scale experiments, or experiments that use very similar RNA input from the same source, a qPCR assay is only needed to establish the optimal cycle number during the setup of the procedure. Further experiments using the same overall sample and input conditions can reliably use this optimal cycle number without the need to repeat the qPCR assay for each new set of library preparations.

4. qPCR as Important Checkpoint for Library Generation from Ultra-low Input RNA and Single Cells

For library preparation from single cells or ultra-low input RNA, performing qPCR during the library generation process is one way to assess the quality and quantity of the input prior to analysing the final library traces. The RNA content from these samples is too low to be detected by any other means, therefore, RNA quality control mechanisms cannot be used.

Single-cell protocols that use PCR to generate amplified cDNA prior to library conversion by a DNA library prep offer the possibility to run the amplified cDNA on a microfluidics instrument for quality control. Many of these protocols additionally recommend cDNA quantification by qPCR to determine the appropriate input amount for the subsequent DNA library preparation.

The qPCR quantification during library generation is also used to determine the optimal number of PCR cycles for library amplification (see below).

5. Why Less is more: Using qPCR Assays for Optimal Cycle Number Determination

As only a very small proportion of library fragments is finally sequenced (see Chapter 2), all sequencing workflows possess an inherent sampling bias. Further, PCR reactions generate more errors and artifacts when a high number of cycles is applied, as the reaction runs out of critical components by depleting the individual nucleotides unevenly and primers becoming scarce (please see Chapter 7 for more details). A higher library concentration is therefore often a trade-off with quality / complexity (low duplication rate) and accuracy (low error rate).

Adaptation of the PCR cycle number is necessary to account for the varying content of amplifiable target sequences in the samples, e.g., when working with samples from different sources or samples of heterogenous quality, such as intact and degraded or FFPE samples. Choosing an incorrect cycle number can either lead to insufficient amplification (termed undercycling) or to excessive amplification (termed overcycling).

Undercycling generates libraries with yields that are too low for accurate quantification, size estimation, or lane mixing. The yield of these libraries can be increased by adding additional PCR cycles – however, this uses additional resources and can introduce further bias and loss in complexity. Overcycling is characterized by formation of aberrant products due to exhaustion of reaction components, which impact accurate quantification, and reduce data quality.

What happens during Overcycling?

Overcycling occurs during the plateau phase of PCR. Reaction components and primers become scarce and heteroduplex products are generated. Each library molecule in the reaction is tagged with the Illumina-compatible adapter sequences. While the insert sequences are variable, the adapters on one molecule are complementary to the adapter sequences flanking any other molecule in the reaction. As a result, adapter sequences can anneal between different library molecules and thus form partially complementary heteroduplex structures. These so called "bubble products" contain double-stranded regions on either end made up by the Illumina-adapters and a non-complementary partially single-stranded bulge corresponding to the insert fragment, the "bubble" (Fig. 5).

During the exponential phase of PCR, heteroduplex formation is prevented, as amplification primers block the adapter regions and mis-annealing to another library molecule is thwarted (Fig. 5).

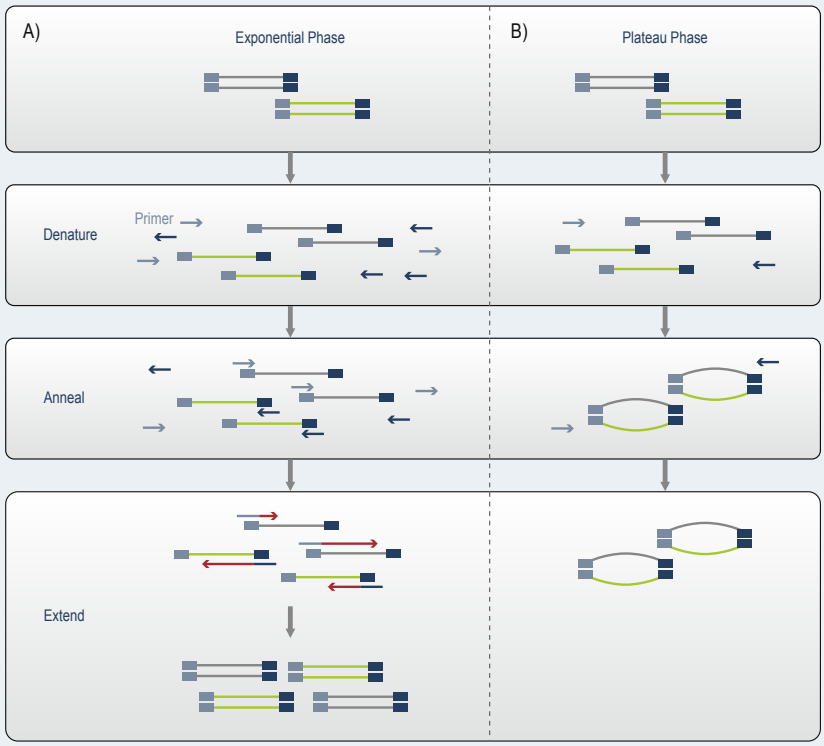


Figure 5 | Formation of aberrant PCR products. A) PCR during exponential amplification phase: double-stranded DNA libraries are denatured, primers are annealed and extended to complete the complementary strand. Then, a new cycle is started. B) When primers are depleted, complementary adapter sequences can anneal to each other generating bubble products. Adapted from Illumina.

When aberrant bubble products occur as a consequence of overcycling, a characteristic high molecular weight "bump" becomes visible when analyzing the affected library traces (Fig. 6). While these libraries are still sequenceable, quantification is impaired often causing unequal read distribution between samples.

Overamplification leads to higher duplication rates, potentially reduced complexity, and higher sampling variance. In the worst case, the data obtained from a sequencing experiment can be distorted. The variability between replicates can thus be increased so much that the data set loses its usefulness. Correct data interpretation can be extremely challenging and dominated by large effects that are brought about by PCR artifacts – in the worst case, leading to incorrect biological conclusions. For more details also see our blog article on library amplification and cycle number determination. By using an inexpensive qPCR assay to determine the optimal PCR cycle number, researchers can avoid these pitfalls and ensure the best results for their sequencing experiments.

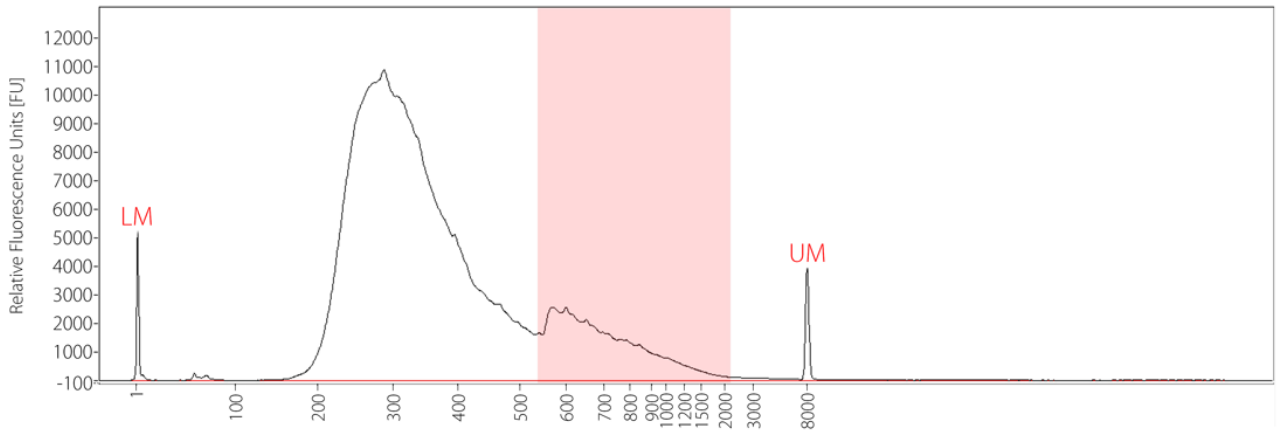
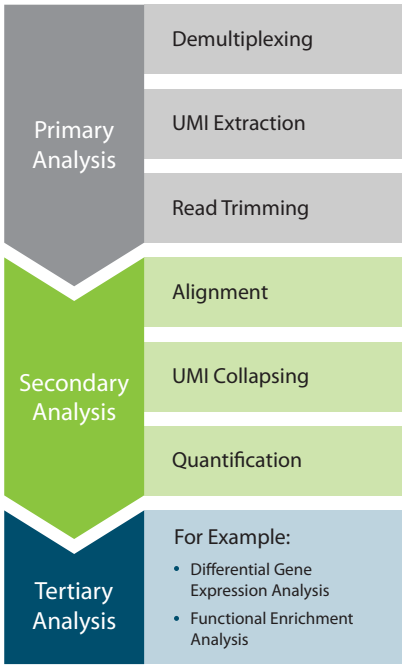


Figure 6 | Fragment Analyzer trace for an overcycled library showing a characteristic bubble product.

Data Analysis and Quality Control – Primary Analysis

The increase in throughput and popularity of RNA-seq has resulted in an unprecedented need for expertise in bioinformatics and computational resources. Although many steps in RNA-seq analysis have become standardized, others still pose major bottlenecks or pitfalls. As RNA-seq, the de facto standard of transcriptome profiling, continues to push the boundaries of data output, the demands and expectations of processing this data has increased significantly. This is especially true when used for diagnostic or screening purposes where data analysis workflows must be quality controlled to ensure highly reproducible findings. This series of Lexicon chapters outlines the central data processing and quality control steps essential to RNA-seq analysis.

How is RNA-seq data analyzed?



RNA-seq analysis is commonly divided into primary, secondary, and tertiary analysis (Fig. 1). Primary data analysis includes processing the raw sequencing data. This step consists of demultiplexing the samples according to their respective indices (barcodes) and read trimming. Secondary analysis describes the process of aligning and quantifying the pre-processed reads. Tertiary analysis focuses on extracting biologically relevant information from the samples. Often, this step includes differential gene expression (DGE) analysis and gene ontology (GO) term or gene set enrichment analysis, which are described in our upcoming Chapter 13. As tertiary analysis is the final and most extensive analysis step, it also encompasses visualizing the results. Distilling large amounts of data into comprehensive and easily understood figures is an art of its own. However, data visualization and figure creation are beyond the scope of this article and will not be discussed. Data visualization remains one of the most challenging but exciting aspects of RNA-seq analysis and is a process each researcher should further explore on their own.

Figure 1 | Data analysis steps overview. Primary analysis refers to the initial analysis steps in which the reads are prepared for further processing steps. Secondary analysis which we will focus on in Chapter 12 encompasses alignments, UMI analysis and gene / transcript quantification. Tertiary analysis uses the output generated from the primary and secondary steps to analyze the generated data in a physiological context, examples include differential expression analysis comparing multiple conditions or identification of signaling pathways, regulated targets, or interaction partners.

1. Primary Analysis

In this chapter of RNA Lexicon we will focus on pre-processing of the reads during the primary data analysis steps.

Before the Analysis: Sequencing Run Quality Control

Before starting data analysis, sequencing run performance should be evaluated by assessing a set of parameters that are specified for the individual instruments and sequencing modes. These metrics can be analyzed on the sequencers themselves or by using tools like Illumina's Sequencing Analysis viewer. Typically, the total output of the sequencing run is analyzed as well as quality scores. The overall quality score (Q30) is a measure of the run quality, it is defined as threshold for the percentage of bases that should be called with a quality score of 30 or higher.

The Q-score or per base sequencing quality score is a measure of the probability to call a base incorrectly whereby higher Q-scores indicate a lower probability for incorrect base calling. A Q-score of 30, (Q30) indicates a base calling accuracy of 99.9 %. Lower Q-scores can result in a significant percentage of unusable reads and may result in inaccurate conclusions due to a reduction in base calling accuracy.

Further, cluster densities and reads passing filter (PF) can be analyzed. This internal "chastity filter" is passed in the first 25 cycles and serves to removes unreliable clusters from the image analysis results.

For example, for a NextSeq500 run in high output paired-end 75 mode, Illumina specifications state that 80 % of bases should have a quality score of ≥ 30 (80 % ≥ Q30). The expected data output should be between 50 – 60 giga bases (Gb) at cluster densities between 129 and 165 k/mm^2 clusters passing filter.

All of these metrics should be analyzed and kept within Illumina specifications for optimal sequencing results as over- and under-clustering during the sequencing run can decrease the data quality. For more information, please refer to the Illumina website and specifications.

Base calling and demultiplexing

During sequencing, Illumina instruments generate raw data files in binary base call (BCL) format. The advantage of using this format is that each base is recorded in the exact moment when it is called ensuring efficient data processing by the sequencer. These BCL files are most commonly converted into FASTQ files for downstream analysis. BCL to FASTQ conversion is achieved using Illumina's proprietary software, bcl2fastq. As many samples are usually multiplexed in a single run and sequenced simultaneously, the data needs to be sorted again to distinguish the samples it originated from in a process called demultiplexing (see Chapter 9 for details on multiplex sequencing). The bcl2fastq software therefore demultiplexes reads into FASTQ files based on sample indices. Optionally, it is possible to attempt to correct index sequence errors during the demultiplexing step. Several alternative tools are available for demultiplexing and index error correction, including Lexogen's iDemux which can maximize sequencing output in combination with a sophisticated Unique Dual Index Set.

For a single-read sequencing run, one FASTQ file per sample is produced. Two FASTQ files per sample are created for a paired-end sequencing run, one file for read 1 and another file for read 2. Because FASTQ files are text-based files, they share some resemblance to FASTA files. However, FASTQ files consist of a four lines-per-sequence format, while FASTA files contain two lines-per sequence. These two additional lines contain information, such as quality scores, which describe the statistical certainty of a specific basecall. The FASTQ format is widely accepted as the standard format for storing unaligned NGS reads and can be used as input for a wide variety of primary and secondary data analysis tools (learn more about the FASTQ format by checking the official Format Specifications). Typically, these files are compressed and stored with the file extension *.fastq.gz.

Demultiplexing and Index Error Correction

Due to the immense data output next generation sequencing produces, various samples are usually mixed in a process called multiplexing and then sequenced as a pool. Each sample is barcoded via short, defined sequences that uniquely identify a given sample. Demultiplexing refers to the process of bioinformatically reversing this pooling step. During this process sequencing reads are associated to the samples they originated from based on these index (barcode) sequence tags and sorted into individual files.

Index sequence errors that have occurred during the sequencing workflow can be corrected when the respective indexing strategy was chosen.

Dual index sequencing offers the best chance to identify errors in the index sequence and salvage the reads for later analysis. Once identified, index sequence errors can be corrected. Sophisticated index sequence designs allow identification and correction of more errors and thus can make more reads accessible for further analysis thereby increasing sequencing data output.

Explore Chapter 9 to read more about the principles of index sequence design. Several tools are available for demultiplexing with and without error correction, e.g., Illumina's bcl2fastq software can also perform the demultiplexing step.

Lexogen has generated an alternative tool, iDemux which is freely available on github. iDemux can demultiplex indices in the i7 and i5 position as well as i1 inline indices that are part of the read. The program was originally designed to demultiplex Quantseq-Pool libraries which can be triple-indexed and thus contain all three index types. By allowing for simultaneous demultiplexing and error correction of all indices, the tool saves valuable processing time and can maximize data output by rescuing reads with index errors.

While error correction performs best with Lexogen UDIs, the tool is highly flexible and can be used for demultiplexing of any index and is also compatible with barcode sequences from other vendors.

UMI extraction

When Unique Molecular Identifiers (UMIs) are incorporated into samples during library preparation, their sequences must be extracted from the FASTQ reads bioinformatically (see also Chapter 8 on UMIs). Failing to remove UMIs from sequencing reads can significantly reduce alignment rates when mapping against a reference genome thus increasing the number of potential mismatches. Typically, this is achieved by "splicing out" the UMI sequence from the read (Fig. 2). Subsequently, the UMI sequence is added into the header of the read. This method retains the UMI sequence of each read without interfering with alignment.

Most bioinformatic tools with UMI functionality are designed to simultaneously extract UMI sequences from FASTQ reads and collapse PCR duplicates post alignment. Commonly, collapsing of UMIs relies on positional information gathered during read mapping, i.e., the mapping coordinates of the read associated with the UMI. We will return to UMI collapsing in Chapter 12 when we focus on secondary data analysis.

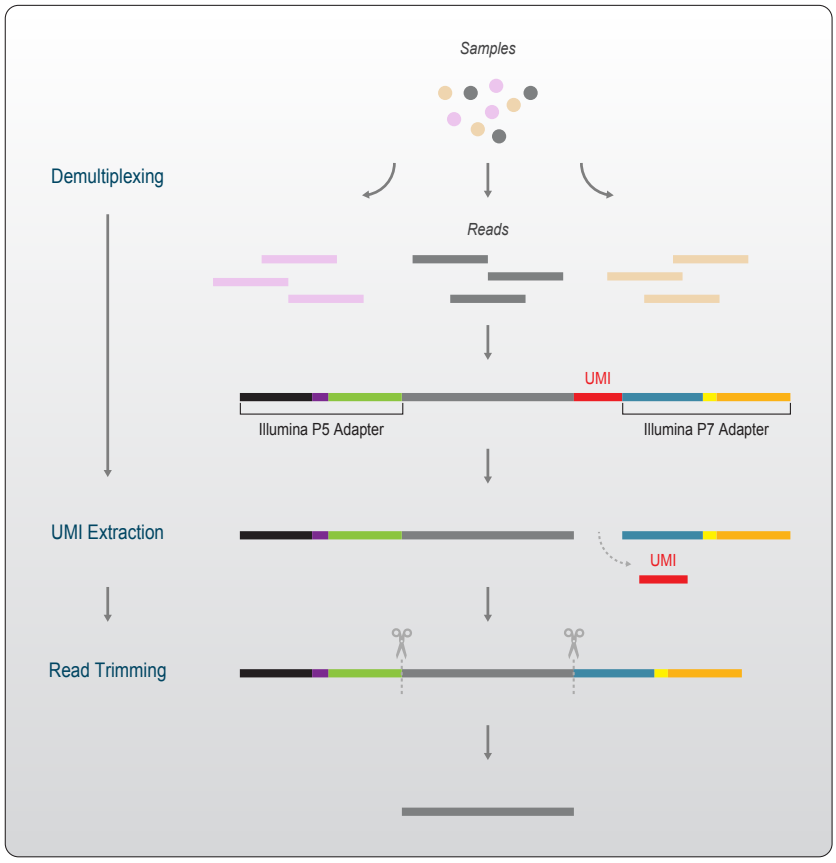


Figure 2 | Primary data analysis. Following the demultiplexing step, UMI sequences are extracted from the read and written into the FASTQ header. UMIs can be located at either end of the sequencing read and their positions is typically specified in the command used for UMI extraction. In the next step, adapter sequences and sequences of low quality, such as homopolymer stretches are trimmed. Failing to remove the UMI or adapter sequences negatively influences read mapping and can lead to low alignment rates.

Read trimming

Prior to mapping reads against a reference genome, it is recommended that the user performs read trimming. Often, Next Generation Sequencing (NGS) reads contain undesirable adapter contamination, poly(A), or poor-quality sequences which should be removed. Failing to remove these problematic sequences may result in reduced alignment rates or false alignments. When utilizing an Illumina sequencer with 2-channel chemistry, it is also advisable to trim poly(G) sequences. These poly(G) sequences result from an absence of signal and will default to G (Fig. 3).

Two popular tools for read trimming are cutadapt[1] and Trimmomatic[2] .

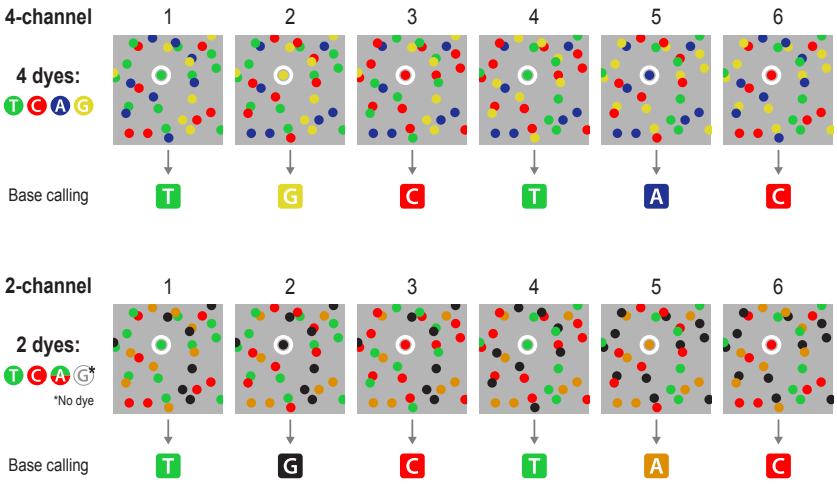


Figure 3 | Base calling for Illumina sequencers using 4-channel chemistry (top) or 2-channel chemistry (bottom). While all four nucleotides are labeled with a fluorescent dye when 4-channel chemistry is used, only two dyes are used to distinguish the four nucleotides when 2-channel chemistry is used. Here, T is labeled green, C is labeled red, and A is labeled both green and red. As red and green overlay for A, a signal can be detected in both channels clearly identifying the base. In contrast, G is unlabeled and therefore, does not generate a signal in any of the channels. Sequences of low quality that fail to generate a signal and remain dark will be called as "G" per default. Poly(G) sequences are therefore often seen in sequencing data from instruments using 2-channel chemistry. Trimming of these sequences improves the quality of the data for subsequent analysis steps.

2. Quality Control

When analyzing sequencing data, it is essential that issues are detected early on. Detecting errors prior to analysis saves valuable time and resources and ensures sound biological conclusions are made. Concordant to the concept "garbage in, garbage out"(Fig. 4), one cannot expect to generate biological meaningful results in tertiary analysis when processing fundamentally flawed data during primary and secondary analysis.

Therefore, strict quality control at each analysis step must be performed to thoroughly understand the strengths and weaknesses of a data set and ensure conclusions are made in good scientific practice.

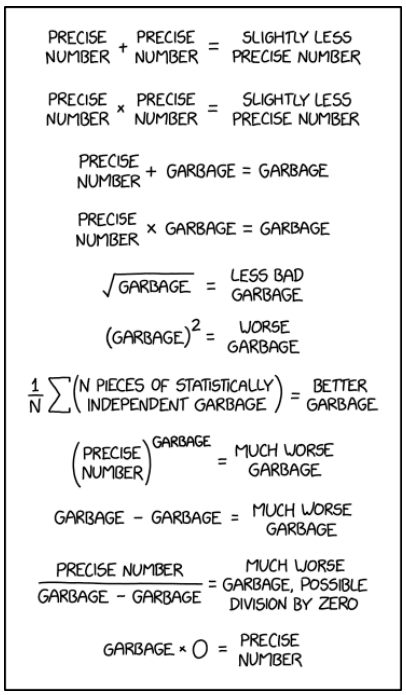


Figure 4 | Input / Output Concept. The quality of the data and control of the processing steps is fundamental for drawing sound conclusions. As the "garbage in, garbage out" concept signifies, flawed data will lead to flawed conclusions. Image curtesy of xkcd comics.

Quality controlling FASTQ reads with FastQC

The most commonly used tool to quality control FASTQ reads is FastQC[3]. FastQC presents the user with a report containing a variety of relevant summary statistics. Each statistic on the FastQC report is scored according to a 'traffic light' system. Green indicates normal data, orange is borderline or slightly abnormal, and red represents unusual or poor-quality data.

At Lexogen, FastQC is run before and after each primary analysis step to determine the impact of the analysis step performed, as well as the effect that data quality will have on downstream analysis. While acceptable and often unavoidable to have some yellow or red warnings during or at the end of primary analysis, the workflow should be optimized to maximize the number of "green lights". Due to the unique library preparation chemistry inherent to Lexogen products, certain FastQC summary statistics may be flagged as unusual in the report. However, this is normal and expected.

For example, when using a library preparation protocol with UMIs and running FastQC before trimming or UMI extraction, the "Per Base Sequence Content" will typically display a red light. Adapter or UMI sequences naturally bias the "Per Base Sequence Content" and will be flagged as unexpected. Products from the QuantSeq family may also display a warning for the "Sequence Duplication Levels" statistic. This is because FastQC was originally designed as a quality control method for whole-genome shotgun sequencing data where high sequence diversity is key. For 3' transcriptome sequencing methods such as QuantSeq, reads are concentrated at the 3' UTR, which results in an overall lower sequence diversity and overestimation of potential PCR duplicates.

Software specific diagnostic output

In addition to examining FastQC results, it is also advisable to study the diagnostic output of the different bioinformatic tools used.

For example, Bcl2fastq will output the number of reads demultiplexed for each sample. Unexpected ratios, or a very low number of reads for all samples can indicate incorrect sample barcoding or that the wrong barcodes have been supplied to the software. Similarly, trimming software such as Cutadapt or Trimmomatic also provides diagnostic output on the percentage of total reads and bases trimmed, as well as the average length of the remaining reads.

After successful pre-processing and read quality control, the reads are prepared for the next analysis steps. Stay tuned and read up on Secondary Data Analysis in Chapter 12 of our RNA Lexicon.

Literature:

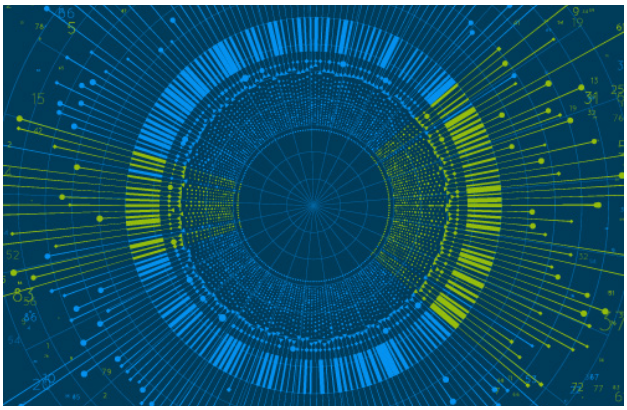
1. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17: 10-12. DOI: 10.14806/ej.17.1.200

2. Bolger AM, Lohse M, Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 30:2114-2120. DOI: 10.1093/bioinformatics/btu170, PMID: 24695404; PMCID: PMC4103590.

3. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Data Analysis and Quality Control – Secondary Analysis

Next Generation Sequencing (NGS) technologies offer high-throughput, rapid, and accurate methods to assess genomes and transcriptomes in all life science fields. Correspondingly, the need for bioinformatic assessment of biological data is continuously increasing. Bioinformatics tools organize, analyze, and interpret biological information at molecular or physiological level to drive basic and applied research. In this chapter of RNA Lexicon, we will continue to explore central data processing and quality control steps used for RNA-Seq analysis. Following the pre-processing steps outlined in Chapter 11, this chapter focusses on secondary data analysis which provides the fundamental basis for assessing any research question tackled by NGS approaches.



1. Alignment

Annotation-based Aligners

During alignment, sequencing reads are mapped against a chosen reference genome. This alignment produces a Sequence Alignment Map (SAM) file, or its binary and compressed counterpart, a BAM file (for a closer look at the SAM file format, check out the SAM format specifications). These files contain the genomic mapping locations of all reads, in addition to a CIGAR string. The CIGAR string is a sequence of base lengths and an associated operation that describes the mapping quality and the length of potential gaps and insertions. An example for the information contained in a CIGAR string is given in Fig. 1.

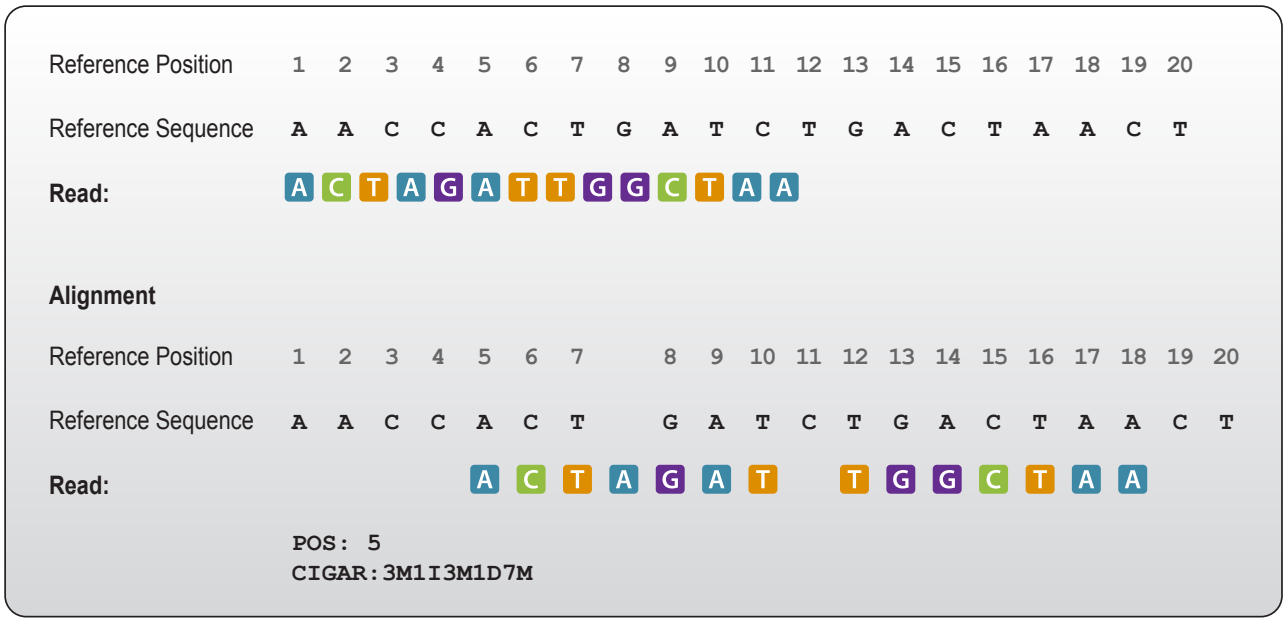


Figure 1 | Example of a CIGAR string for alignments. "POS" defines the position in the reference at which the read alignment starts, in this case, position 5. The CIGAR string states "3M" therefore 3 bases of the read map to the reference. "1I" defines one base that is inserted in the read but does not exist in the reference. This insertion is followed again by 3 bases that map to the alignment indicated by "3M". Next, one base missing in the read (but present in the reference) is referred to as 1 deleted base "1D". The last part "7M" defines seven bases that are aligned to the reference. Note that the base on position 14 is a mismatch to the reference, it is counted as mapped as it occupies a position rather than generating an insertion or a gap.

In other types of alignments CIGAR strings also specify clipped bases, mismatches, or longer gaps in case of spliced introns. These alignments per gene or transcript can later be counted to determine their expression. Popular aligners utilized for RNA-seq are TopHat[1], HISAT[2] and STAR[3]. All these aligners can perform gapped alignment. Gapped alignment is essential for RNA-seq, as splicing of transcripts may lead to large alignment gaps which need to be accounted for (Fig. 2).

The genomic reference contains the complete sequence space of an organism including sequences that are not transcribed (e.g., intergenic sequences without overlaying transcripts) or sequences that are spliced out after transcription (e.g., intron sequences). As RNA-Seq predominantly aims to capture mature transcripts, most of those sequences are also absent from the reads generated in an RNA-Seq experiment. Therefore, two sequences that are next to each other in a read can originate from loci that are much further apart in the genomic reference. Aligners thus need to identify these sequence parts or "seeds" and account for the gap in the alignment (Fig. 2B).

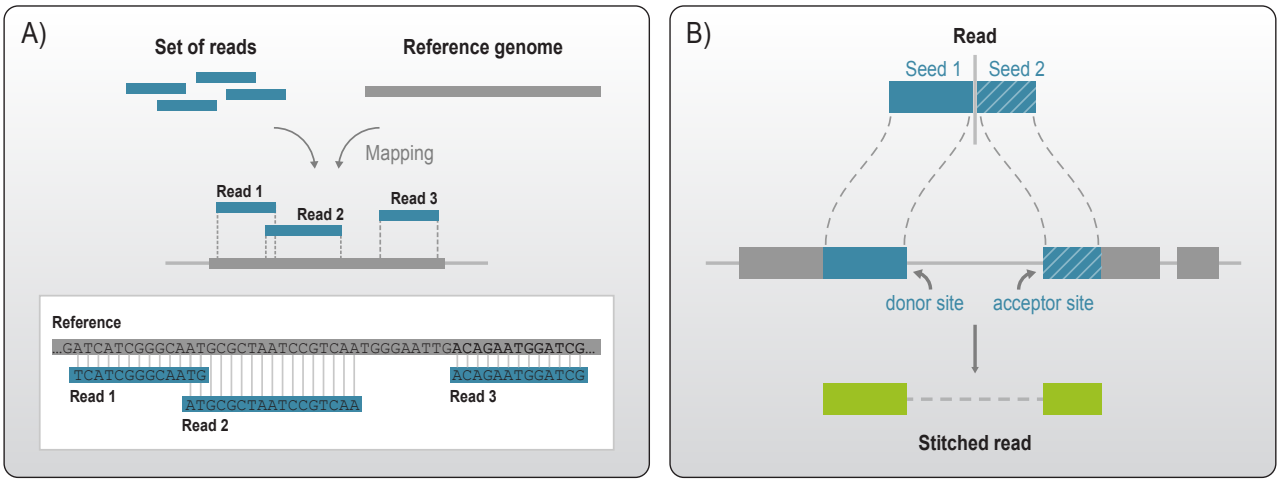


Figure 2 | Read alignment with annotation-based aligners. ***A)*** *The set of sequencing reads is mapped against a chosen reference genome based on sequence similarity.* ***B)*** *Gapped alignment of RNA-seq reads. When mapping RNA-Seq reads to genomic sequences, aligners need to operate by accounting for gaps. RNA-seq aligners as for instance STAR operate by finding the longest sequence in a read matching the reference genome (seed1). If the read cannot be mapped completely to one continuous sequence, STAR will then try to find the mapping location of the remaining sequence (seed2). The separate seeds are then stitched together to get the alignment position of the read. If this splice junction is not yet known in the reference genome, STAR will also report it as a new junction. Adapted from [3].*

Pseudo-aligners

Recently, a special class of aligners, referred to as 'pseudo-aligners', have been developed to drastically decrease alignment time. Popular pseudo-aligners like Salmon[4], Kalisto[5], and Sailfish[6] can decrease runtime up to 250-fold, enabling users to quantify their sequenced reads on a simple desktop computer in approximately 10 minutes. Pseudo-aligners utilize sophisticated statistical algorithms to assign a read to a given transcript sequence without mapping the read to the actual genomic location. First, the sequences of all known reference transcripts are broken down into so called k-mers, short subsequences of roughly 30 base pairs. The pseudo-aligner then constructs a network, where these short subsequences serve as nodes. These nodes are then connected by different paths that describe the sequence of a transcript (Fig. 3). The pseudo-aligner then estimates the most likely path (transcript) the sequencing read originates from [5]. As a result, pseudo-aligners do usually not output SAM or BAM files, but only tab-delimited quantification tables (although Salmon and Kallisto nowadays can be set to output other file types as well). Therefore, this method reduces the transparency of the quantification process and does not allow the user to look at the read distribution at specific loci. This lack of read distribution may be detrimental when attempting to answer specific research questions. Most pseudo-aligners are also unable to perform read collapsing based on Unique Molecular Identifiers (UMIs) for bulk RNA-seq data analysis.

Which Aligner is Best used for your Analysis?

Due to the limitations outlined above, it is recommended to use annotation-based aligners to analyze RNA-Seq data from libraries containing UMIs (e.g., data derived from CORALL library preps) when UMI-based read collapsing is required. Annotation-based aligners, specifically STAR is also recommended when using 3' mRNA-seq methods (such as Quantseq), although benefits of using pseudo-aligners were highlighted especially for high-throughput 3'-Seq projects[7].

The downside of using annotation-based aligners is that they require a high-quality genome annotation to operate properly. While annotations for model organisms, such as human and mouse genomes, are continuously updated, researchers can be hard pressed to find a decent annotation when working with non-model organisms. In case of 3'-seq data analysis, a decent 3' annotation is especially important to maximize the percentage of mapped reads for a meaningful analysis. In this case, pseudo-aligners allow to analyze and quantify expression of transcription units independent of an annotation and thus can offer more information to the researcher.

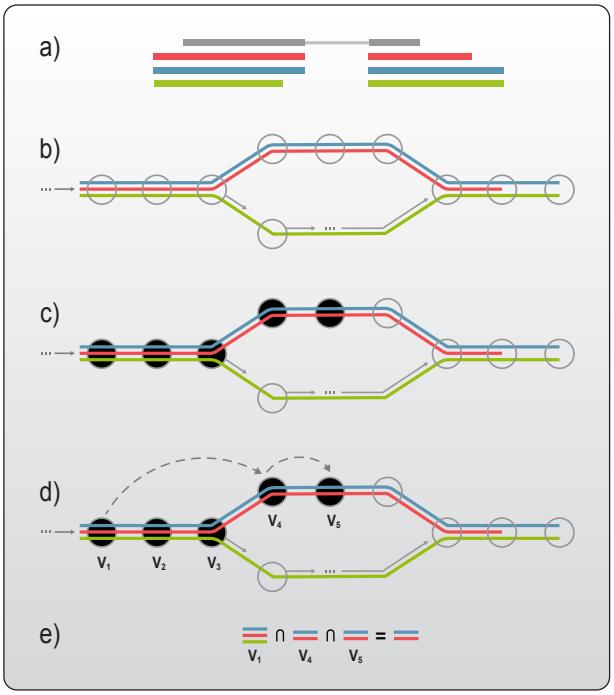


Figure 3 | Read alignment with pseudo-aligners. When using pseudo-aligners such as Kallisto, reference transcripts are broken down into much smaller subsequences (k-mers). These subsequences are then used to construct a network, where each subsequence is a node. In this network, transcripts can be described by a path leading from node to node. Reads that should be pseudo-aligned are then split as well into multiple k-mers, while the algorithm calculates which path (transcript) the deconstructed reads most likely originates from. Figure adapted from [5].

2. UMI-based Read Collapsing

If UMIs have been utilized during library preparation, UMI-based read collapsing can be performed. In general, this analysis step is optional and UMI-based deduplication can be omitted when working with high input amounts to save computational resources. In any case, the UMI sequence should always be removed before proceeding to the alignment step to ensure unimpaired read mapping (see Chapter 11 for details).

UMI collapsing serves to remove PCR duplicates by reducing them to one read. During PCR, DNA fragments amplify with varying efficiencies. How efficient a fragment is amplified is often based on multiple factors such as length, GC content, secondary structure and the number of PCR cycles that are applied to an NGS library (see Chapter 7 for more details). UMI-based read collapsing can reduce the biases which are introduced during amplification.

While the "duplication rate" assessed by popular sequencing quality control tools such as FastQC can be seen as an indication for the complexity of a data set, UMI deduplication offers a more precise estimation of actual read duplication during the complete sequencing workflow including the PCR step and sequencing itself.

Most bioinformatic tools achieve this by comparing the UMI sequence of reads with the same starting position. Reads with the same UMI are marked as PCR duplicates and collapsed into one single read (see Chapter 8 for a detailed overview of UMIs). The most widely used and recommended software for UMI-based read collapsing is UMI-tools[8]. Commonly used tools for UMI-based collapsing also offer the possibility to correct for errors within the UMI sequences (Fig. 4).

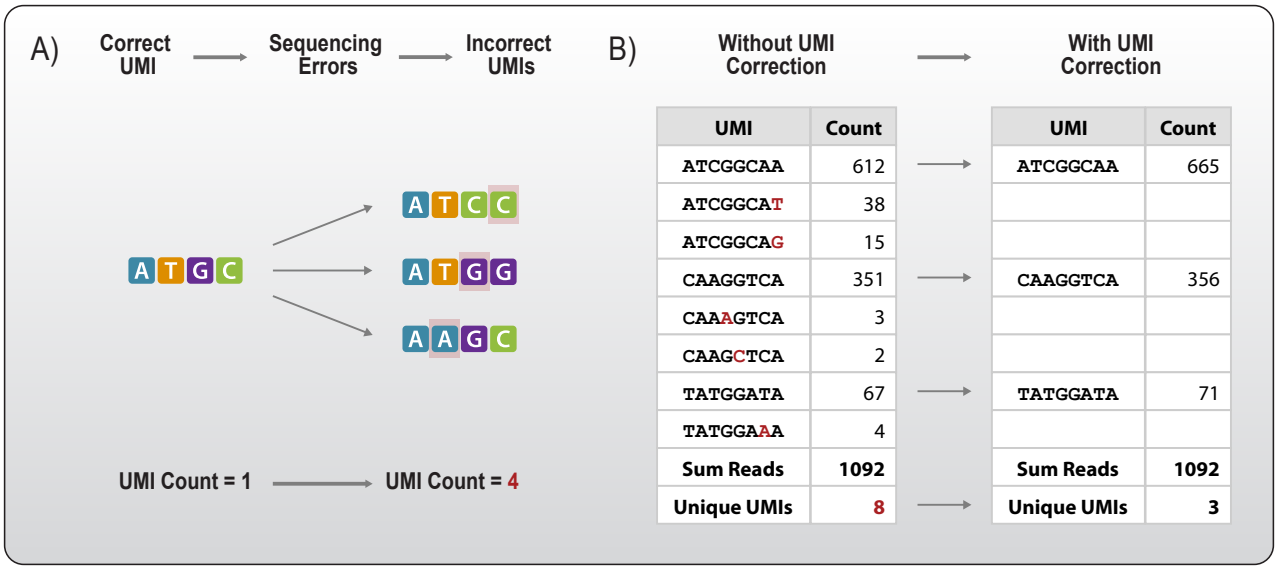


Figure 4 | UMI error correction. ***A)*** *Sequencing errors introduced into the UMI sequence lead to novel sequences that differ in a limited number of bases, in most cases 1 error is set as default. Counting UMI sequences without error correction would result in a unique UMI count of 4 instead of 1.* ***B)*** *UMI counting with and without error correction. UMIs mapping to the same gene / position are counted to improve quantification of the gene or transcript (see Chapter 8). Counting UMIs that contain sequencing errors without correction overestimates the number of unique UMIs and can thus impact quantification. During the correction step, the UMI with the highest number of assigned reads is assumed as correct or parent sequence. UMIs that differ in 1 base have a Hamming Distance of 1 and are likely derived from the parent UMI (see Chapter 9 for more information on sequence distances). Thus, UMI error correction will group these UMI sequences reducing the count of unique UMIs for the respective gene / position.*

UMI-tools does this by building a nearest neighbor network of UMI sequences for all reads at a specific mapping location. UMIs that can be derived from each other with a substitution of one base (or a higher user defined threshold) will be connected and grouped together in one network. This network is then used to identify potential sequencing errors in the UMI based on the network connectivity and collapse the corresponding reads. Depending on the research question, error correction during UMI collapsing can improve the accuracy of the data.

The percentage of reads that are collapsed based on their UMI during this step largely depends on the experimental setup:

- ✓ **Input RNA amount:** processing of limited RNA inputs often requires more PCR cycles for amplification and generates libraries with lower complexity. As a consequence, a large number of reads can be collapsed during this step, especially when sequencing at saturated depth.
- ✓ **Complexity of the sample:** higher collapsing rates are seen for samples which are low in complexity due to a lower diversity.
- ✓ **Sequencing depth:** at saturated sequencing depth, a larger fraction of reads is expected to be removed by collapsing. At which per-sample-sequencing-depth saturation is reached depends on the complexity of the sample itself, the input amount, and the efficiency of the library preparation. For samples with low complexity, at low input amounts or for library preps with low efficiency, saturation is reached earlier at lower depth per sample.

For example, sample types with a lower complexity or limited input material are expected to show higher fractions of collapsed reads, especially when sequencing has reached saturation, i.e., when higher sequencing depth is applied than needed (over-sequencing). As over-sequencing provides little gain in relation to the additional cost, UMIs are very useful to determine the "sweet-spot" of per-sample-sequencing-depth, especially for large scale projects. UMI collapsing rates at different sequencing depths (e.g., using sub-sampling of reads) can thus be used to determine how many reads per sample are optimal for the research question under investigation. This allows to save sequencing costs by avoiding over-sequencing and by multiplexing more samples per run at optimal depth.

3. Quantification

During quantification, reads mapping to specific genes or transcripts are counted to provide a direct readout of gene expression. Many popular quantification tools are available, such as featureCounts[9], RSEM[10] and Salmon[4]. Lexogen has also developed a unique proprietary quantifier, Mix^2 [11], which quickly and accurately estimates transcript concentrations.

Most bioinformatic tools for differential gene expression, a popular tertiary analysis, use raw counts as input. While raw counts are the output of all quantifiers, many also provide differently normalized expression values. Hence, quantification tools have different specializations. Therefore, it is important to consider the research question at hand and library type when selecting a quantifier. The inputs for most quantifiers are usually a SAM or BAM file, which contains the aligned reads and a genome annotation. The quantifier then determines the number of reads overlapping with annotated transcripts or genes. FeatureCounts, one of the most minimalistic quantifiers available, returns a count table for each gene or transcript making it an especially useful tool for 3′ RNA-Seq methods where reads are localized to the 3′ ends of transcripts. Thus, 3′ RNA-Seq data sets do not require fitting of sequencing reads over the complete transcript length and can also forego transcript quantification estimation.

For example, due to QuantSeq's straightforward one-read-per-transcript chemistry and the resulting simplicity of analysis, featureCounts is recommended for quantifying QuantSeq data.

In contrast, limitations apply when counting is used on whole transcriptome (WTS) data sets. As longer transcripts naturally receive more counts than shorter ones, counting introduces a length bias leading to a relative inaccuracy between different transcript length classes in one sample. Therefore, these data sets are commonly analyzed with tools allowing for transcript concentration estimation, such as RSEM, Salmon or Lexogen's Mix^2. Next to raw counts and length all three of these algorithms return normalized expression values such as FPKM (Fragments per kilobase of exon per million fragments mapped) or TPM (Transcripts per million). This allows the researcher to compare expression values of genes of varying lengths within one sample.

Useful Tips for Data Analysis of Lexogen Libraries

Data analysis for Lexogen libraries differs slightly from the commonly used settings for analysis of data generated from other vendor's library preps. **Most of the library preps developed by Lexogen generate reads in forward orientation, i.e., Read 1 reflects the sequence of the RNA transcript (Fig. 5)**.

In contrast, many conventional library preps generate reads in reverse orientation, i.e., Read 1 reflects the sequence of the cDNA, the reverse complement of the RNA transcript.

Researcher therefore need to take care when using freely available tools for data analysis as standard settings may reflect reverse read orientation. Performing data analysis steps with incorrect read orientation is known to produce false results, for example for genome wide coverage analysis. Gene body coverage plots are calculated from the coverage distribution across a set of abundant genes along their normalized length. As they are often used as a quality parameter, a false read orientation setting will give the impression of a poor-quality data set. Further, gene and transcript quantification will be affected if the read orientation is set incorrectly.

UMIs can also often be difficult for researchers to process. In general, the position of UMIs is variable and depends on the respective library prep itself. Most UMIs are positioned at the beginning of Read 2 and require (partial) paired-end read mode during sequencing, as for example in QuantSeq-Pool libraries (Fig. 5). In contrast, for QuantSeq FWD and CORALL, the UMI is located at the beginning of Read 1 and is thus naturally a part of Read 1 (Fig. 5). As such, the UMI needs to be extracted prior to the alignment step to avoid interference during mapping as described in Chapter 11.

To facilitate data analysis for our users, Lexogen offers various plug-and-play data analysis solutions at partner platforms as well as a set of pipelines and tools on Github. For more information on data analysis solutions, visit our online FAQ page.

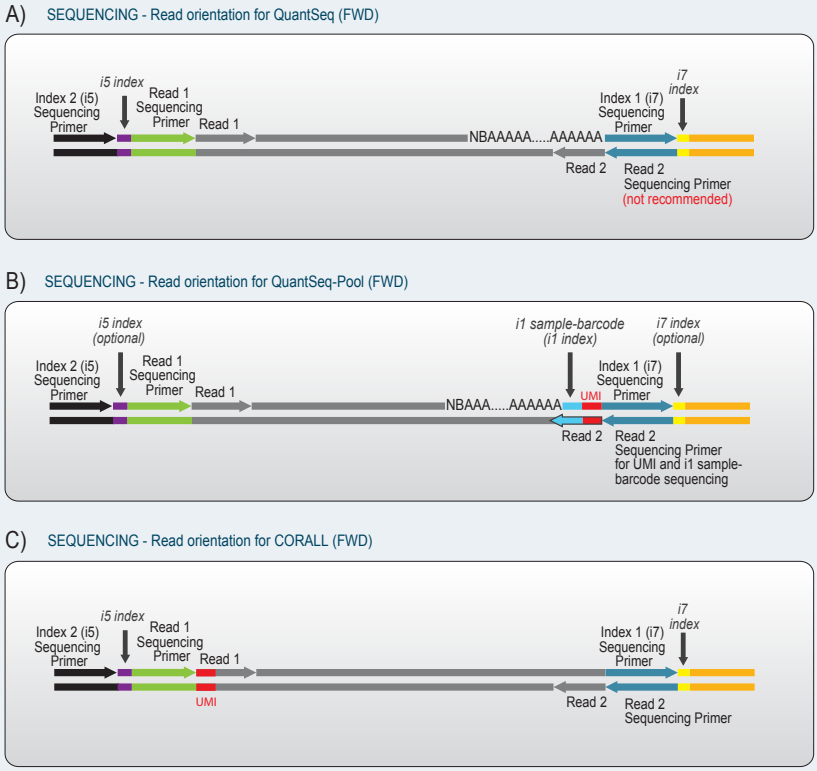


*Figure 5 | Read orientation for **A)** QuantSeq, **B)** QuantSeq-Pool, and **C)** CORALL libraries. For all libraries, Read 1 is in forward orientation and thus corresponds to the transcript sequence. While the UMI and inline index for QuantSeq-Pool libraries is located at the beginning of Read 2, the UMI sequence for CORALL libraries is read out at the beginning of Read 1. Optionally, a UMI can also be introduced in classical QuantSeq libraries. In this case, the UMI is also located at the beginning of Read 1 (not shown).*

The expression information generated during the quantification step is the basis for more advanced tertiary analysis. In our final Chapter on Data Analysis, we will explore the most common applications for tertiary analysis and add further insights on quality control steps that will help you to make the most of your RNA-Seq data.

Literature:

1. Kim, D., Pertea, G., Trapnell, C. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14: R36. doi:10.1186/gb-2013-14-4-r36

2. Kim, D., Langmead, B. and Salzberg, S. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12: 357–360. doi:10.1038/nmeth.3317

3. Dobin, A., Davis, C. A., Schlesinger, F., *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15–21. doi:10.1093/bioinformatics/bts635

4. Patro, R., Duggal, G., Love, M. et al. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14: 417–419. doi:10.1038/nmeth.4197

5. Bray, N., Pimentel, H., Melsted, P. et al. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34:525–527. doi:10.1038/nbt.3519

6. Patro, R., Mount, S. & Kingsford, C. (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* 32:462–464. doi:10.1038/nbt.2862

7. Corley, S.M., Troy, N.M., Bosco, A. et al. (2019) QuantSeq. 3′ Sequencing combined with Salmon provides a fast, reliable approach for high throughput RNA expression analysis. *Sci. Rep.* 9: 18895. doi:10.1038/s41598-019-55434-x

8. Smith, T., Heger, A., Sudbery, I. (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* 27:491-499. doi:10.1101/gr.209601.116

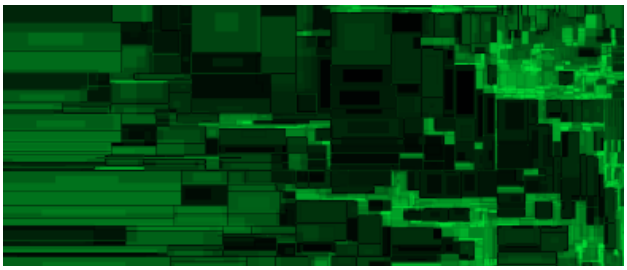
9. Liao Y, Smyth GK, Shi W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* 30: 923-30. doi:10.1093/bioinformatics/btt656

10. Li, B., Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12: 323. doi:10.1186/1471-2105-12-323

11. Tuerk, A., Wiktorin, G., and Güler, S. (2017) Mixture models reveal multiple positional bias types in RNA-Seq data and lead to accurate transcript concentration estimates. *PLoS Comput. Biol.* 13: e1005515. doi:10.1371/journal.pcbi.1005515

Data Analysis and Quality Control – Tertiary Analysis

The last step of data analysis can be generally described as using tools to convert sequencing data into knowledge and setting it into the biological context. In this Chapter of RNA Lexicon, we will focus on the most common tertiary data analysis types. Before entering tertiary analysis, it is advisable to evaluate the results of the previous steps by a set of additional checks. This way, the researcher can ensure that the data that is used as input for the final analysis steps has passed all quality control standards.



1. Quality Control Before Entering Tertiary Data Analysis

Quality controlling the results of secondary analysis data ensures that the subsequent tertiary analysis steps are conducted with high quality data and scientifically sound conclusions can be drawn from the final output. Quality control at this step is centered on verifying that the distribution of reads matches the *a priori* expectation and that the quantification process will provide an accurate read-out of the library input.

Alignment Rates and Read Distribution

Once the data has been aligned to a reference genome, the aligner will output basic summary statistics. These statistics usually include the percentage of reads mapped to the reference genome. For an ideal RNA-seq library, this metric should be greater than or equal to 90 %. While alignment rates close to 70 % may still be acceptable depending on the quality of the RNA input and the reference genome used, lower alignment rates may indicate serious issues with the data set.

Using mapping rates as QC parameter is only possible when working with organisms which are well-annotated. For non-model organisms, genome assemblies and annotations are often poor and / or incomplete. In this case, low mapping rates are to be expected and are mostly caused by the reference rather than the quality of the data set.

One explanation for low mapping rates observed for well-annotated model organisms is that most reads are too short to be properly mapped to the genome. This situation can arise when highly degraded RNA is used as input, the libraries or sequencing run are poor in quality, or when the reads have been trimmed too short in length. Another potential explanation for poor mapping quality is contamination of input material with foreign RNA. Construction of the first tardigrade genome assembly is a classic and well-cited example of how contamination can negatively influence NGS library composition and lead to false conclusions[1, 2]. Bacterial contamination in tardigrade cultures led to an overestimation in the amount of horizontal gene transfer that occurred in this genome.

When low mapping rates are observed, it may be useful to simply BLAST a portion of the unmapped reads to uncover their biological origin. However, when mapping percentages do not indicate any obvious problems, it is useful to visualize read distribution across different genomic features.

For example, RSeQC[3] can be used to determine the percentage of reads which map to the CDS, 5', and 3' UTRs or the intronic or intergenic space. Another software with similar functionality is Picard tools.

Read distribution is an important metric which enables the user to gauge if the library contains expected read fragments. For 3' mRNA-seq library preps such as QuantSeq, most reads should be concentrated at the 3' UTR. In contrast, for whole transcriptome sequencing (WTS) library preps most reads typically map across the complete transcript body (Fig. 1). A concentration of reads towards the 3'UTR would indicate degradation of the RNA sample prior to library generation. The distribution of reads over the whole exonic space or the coding sequence depends on whether upstream rRNA depletion or poly(A) selection was performed, which also has implications on the percentages of intronic and intergenic reads.

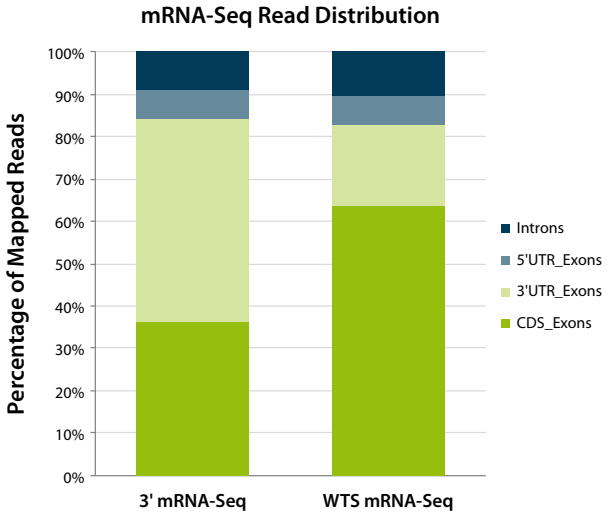


Figure 1 | Mapping class attribution for reads generated using 3' mRNA-Seq or Whole Transcriptome mRNA Sequencing (WTS). Reads generated by 3' mRNA-Seq are located towards the 3' UTR of transcripts, as represented by the majority of mapped reads. In contrast, reads obtained for mRNA WTS-libraries are distributed evenly across the complete transcripts. Therefore, reads mapping to coding sequences should represent the majority of mapped reads and the fraction of reads mapping to 3' UTRs is lower than for 3'-Seq.

For example, data generated from poly(A)-selected RNA typically reflects mature mRNAs with a lower intronic and intergenic read fraction. Due to the possibility to capture pre-mature mRNA, intron-inclusion events, and the quality of the annotation itself, a certain level of intronic and intergenic reads is to be expected, whereby the intronic read percentage should be higher than the intergenic read percentage. For data generated from rRNA-depleted samples more intronic and intergenic reads are expected as this method also captures transcripts occupying this space, e.g., long (intergenic) non-coding RNAs (lncRNAs and lincRNAs). Further, commonly observed read distribution is also influenced by the sample itself. For example, RNA-seq libraries generated from blood samples naturally show a higher distribution of reads over the intronic and intergenic space[4] (Fig. 2).

A high percentage of intronic or intergenic mapping reads for samples types that routinely show lower values can indicate genomic DNA contamination (most common for WTS data, see also Chapter 5 – DNase: To Treat or Not to Treat). Further, for data obtained from 3'-Seq libraries, such statistics can hint to mis-hybridization where oligo(dT) primers are re-directed from the poly(A) tails of mRNAs and prime to A-rich sequences present in rRNA.

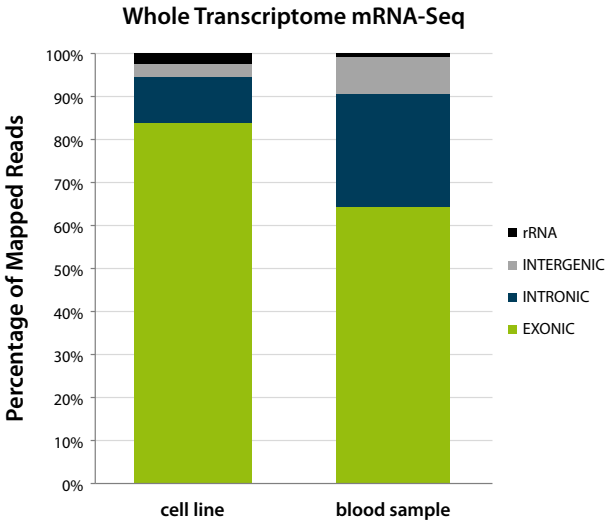


Figure 2 | Feature distribution of mapped reads from Whole Transcriptome mRNA-Seq for different sample types. The majority of reads generated from both samples type, using RNA from cell lines and from blood samples, map to exonic sequences. However, the overall distribution of reads classified as exogenic, intronic, and intergenic is changed depending on the sample type.

Ribosomal RNA as Indicator for Library Complexity

Another important metric to examine is the percentage of ribosomal RNA (rRNA) mapping reads. While total RNA is composed of 80-98 % rRNA, quality mRNA-seq libraries typically contain no more than single digit percentages of rRNA mapping reads.

For example, 3' mRNA-Seq libraries, such as QuantSeq libraries, typically contain ~3-5 % rRNA mapping reads as mitochondrial rRNA transcripts contain poly(A) tails and will be captured by oligo(dT) priming together with polyadenylated mRNAs. In contrast, rRNA depleted WTS libraries, such as CORALL libraries after depletion with RiboCop, typically contain <1 % rRNA mapping reads. The content of reads derived from rRNA observed in sequencing experiments is largely dependent on the sample itself, the RNA quality and quantity, enrichment, and library preparation method. This metric should always be interpreted in relation to expected and typically observed results.

Libraries with a significantly higher fraction of rRNA are usually indicative of low complexity. This can be caused by using low amounts of RNA, or very low-quality input material which usually results in libraires with few detected genes (see Chapter 4 – RNA Extraction and Quality Control). If the genome annotation contains rRNA (some do not), the percentage of rRNA can be calculated from the output of the chosen quantifier. Ribosomal RNA percentage can also be calculated by mapping reads separately to rRNA-only sequences, which can be more accurate when using poor genome assemblies or incomplete annotations. For example, this can be done by mapping the reads to an rRNA-only database such as silva[5] that contains sequences of many different organisms. As rRNA is generally highly conserved this approach can help in these cases to estimate the rRNA content.

Spike-in Controls to Assess Quantification Accuracy and Transcript Coverage

Until now, the Lexicon section on data analysis has primarily focused on read distribution across the genome and how these summary statistics can be utilized as qualitative control. While these statistics provide an adequate overview of the library content and composition, it does not tell the experimenter how accurate the quantification is. If controls such as ERCC[6] spike-ins or Lexogen's Spike-In RNA Variants (SIRVs) are added during library preparation, the researcher can use these as a ground-truth dataset to benchmark quantification performance and detection limits. Further, spike-in controls can be used to fine-tune the entire workflow including data analysis tools and parameters to deliver highly accurate results for the respective research question.

The addition of artificial spike-ins at a low read percentage does not only allow to analyze and compare data sets generated over time and across sites, it also offers the possibility to analyze a small percentage of data for a fast, initial quality control. The spike-in controls are thereby used as a proxy to assess the quality of the library generation and sequencing workflow. Should the sample show unexpected results for any of the parameters outlined above, the artificial controls can help to pinpoint the cause for the observed discrepancies. Internal controls can indicate if there was a sample-related problem, cross-contamination, or difficulties during library generation and sequencing.

2. Tertiary Data Analysis

The tertiary analysis steps depend heavily on the individual research question that was defined at the beginning of the experiment. Therefore, this part of the analysis is the most flexible during the entire project. In the upcoming section, we will therefore focus on some of the commonly used analyses, namely differential expression and functional enrichment analysis. To ensure the success of your project by generating the data you need to answer your specific research question, it is best to consult a bioinformatician prior to starting the experiment (our Chapter 3 on experimental and data analysis planning has you covered and outlines critical considerations for planning your sequencing experiment).

Differentially Gene Expression Analysis

Differential gene expression testing is one of the most common tertiary analysis methods utilized for RNA-seq. Differential gene expression analysis is used to discover significant quantitative gene expression changes under varied biological conditions (Fig. 3 and Fig. 4).

Practical examples for differential expression studies include mapping the transcriptome changes between a *wild-type* and mutant, or expression changes caused by treatment with a specific stimulus or chemical compound, responses to infection, during the course of disease progression or following cell and tissue development.

Two popular tools for differential expression analysis are DESeq2[7] and edgeR[8], both of which operate under the null hypothesis that most genes are not differentially expressed.

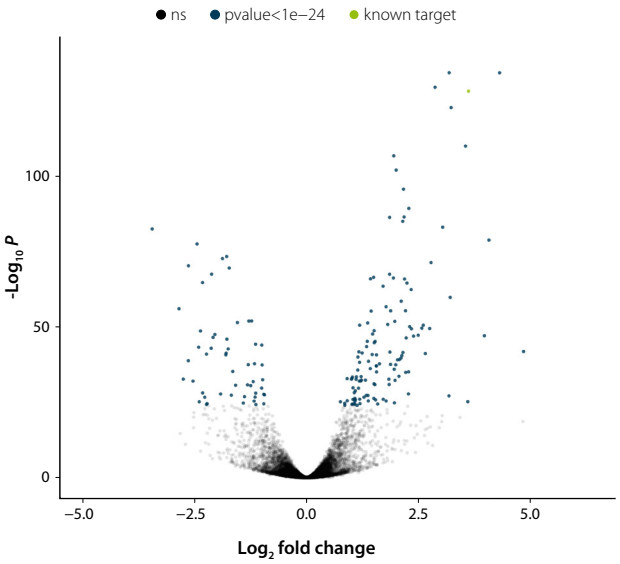


Figure 3 | Volcano Plot to distinguish significant from non-significant changes. Plotting is done based on p-values as measure of significance. Data points with low p-values correspond to highly significant changes and are plotted towards the top. Significant changes are highlighted in blue, known targets in green, and unaffected data points are shown in black. The logarithm of the fold change between the two conditions is shown on the x-axis.

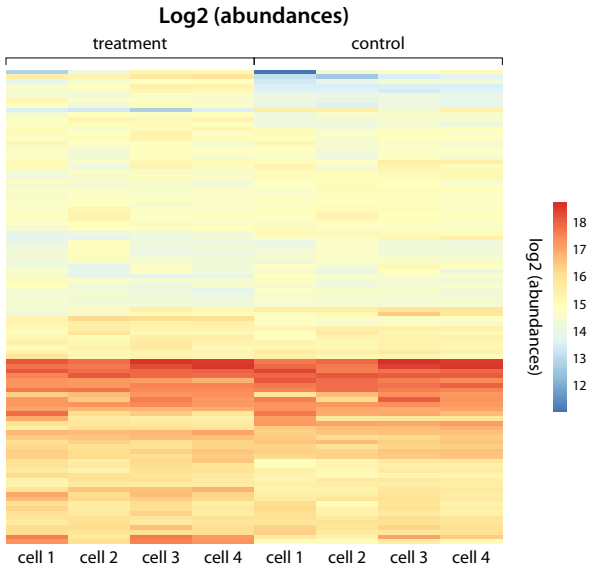


Figure 4 | Heatmap visualizing gene expression changes for various cells under two different conditions. The heat map represents color-coded expression levels of differentially expressed genes, changes in expression level are shown as log2-fold abundances.

The Null Hypothesis and Differential Expression Analysis

Differential expression analysis tools are based on statistical models to estimate the probability if gene expression changes are a result of chance or caused by the varied condition applied. These models operate under the reasonable assumption that most genes are not differentially regulated, and the observed variation is a result of chance. This assumption is also referred to as the *null hypothesis* while the model describing this distribution is called the *null distribution*. During the analysis these tools then calculate how likely it is for each gene that the observed variation is caused by chance, meaning how likely it follows the *null distribution*. This probability is usually expressed as a value between 0-1, the p-value. A value close to 1 indicates a high probability that the observed variation is indeed caused by chance, while a value close to 0 signifies that the *null hypotheses*, the assumption that a gene is not differentially regulated, should be rejected, and the variation is likely causal. Commonly used, but somewhat arbitrary thresholds to indicate significant derivation from the *null hypothesis* are 0.01, 0.05 or 0.1 (DESeq2 default). However, the more tests you perform, and as we are doing this for each expressed gene, there will be a lot, the higher the probability that you will also encounter a low p-value by chance. When performing for exam-

ple 100 tests with a significance threshold of 0.01, you have the probability of encountering 1 significant result just by chance. To account for this multiple testing correction needs to be performed. This can be done for instance by lowering the p-value threshold to 0.0001 (0.01÷100). This approach is called Bonferroni correction and is rarely applied in differential gene expression testing due to its very high stringency. This stringency can limit the discovery of true positive events severely (when testing 10,000 genes this would result in a p-value cutoff of 1×10^{-7}). Therefore, the most common approach is to control the false discovery rate (FDR) via the Benjamini-Hochberg method. In this case the researcher can set the allowed proportion of false positive discoveries (e.g., 0.05 or 0.1) that is acceptable for him. The p-values from each test will then be adjusted based on the likelihood of their FDR. These corrected p-values are thus often called adjusted p-values (padj) or q-values.

The p-values obtained using these tools indicate the probability of a gene not being differentially regulated. Thus, a small p-value leads to a rejection of the null hypothesis and indicates significant differences in gene expression. Both DESeq2 and edgeR statistical models have been designed to work best with raw read counts[6,7]. Gene length normalization is unnecessary as testing for each gene is performed separately and therefore stays constant. As raw read counts do not account for effects such as varied sequencing depths, read counts are normalized in a slightly different manner depending on the tool. For instance, DESeq2 normalizes read counts by multiplying all counts for each sample with a so called "size factor". In concordance with the null hypothesis, these size factors are calculated with the objective to minimize the total variance across each gene for all samples. DESeq2 then estimates the dispersion for each gene (a measure of how much a sample fluctuates around a mean value) followed by a statistical test (for a detailed explanation, we recommend the resources on bioconductor.org, for example The theory behind DESeq 2).

Variations in sequencing depth can be kept low by adjusting the libraries in a lane pool according to their molarity and size distribution prior to sequencing. Equimolar pooling of each individual library enables to sequence each sample with equal read depth during the sequencing run. Even though normalizations can correct variations in read depth, reliable high-quality results are obtained when the variations are already small to begin with. Larger variations can lead to more noise and changes in expression can be harder to detect, especially when the change is rather moderate.

Tools such as DESeq2 and edgeR are also capable of performing tests for classical pairwise comparisons (e.g., control vs. treatment) and more complex scenarios such as time series or effects of a treatment on different genotypes. These complex setups are tested with a likelihood ratio test (to learn more about how these tests are performed, we recommend the resources on Bioconductor.org, for example: likelihood ratio test).

Functional Enrichment Analysis

After differential gene expression analysis has been performed, researchers often desire to gain insight into the cellular functions and molecular processes affected. One way to address this challenge is to annotate genes with metadata which describe their function. This can be achieved by analyzing information on gene-phenotype relationships, associated gene pathways, enzymatic classification of gene products, or organelle function. Once genes have been annotated with this meta data, one can cross-check if genes of a specific pathway are enriched in the differentially regulated genes derived from the RNA seq analysis (Fig. 5). The Gene Ontology (GO) Consortium provides an excellent resource of metadata in the form of standardized terms, intended to represent current scientific knowledge of the functions of genes. GO term annotation and enrichment analysis can be performed online on the Consortiums webpage, or with standalone tools like clusterProfiler[9]. Both options take two lists of gene IDs as input: a background list (non-differentially regulated genes) and a list to test (differentially regulated genes). The standalone tool clusterProfiler, also allows users to incorporate other resources such as KEGG, Reactome, WikiPathways or the molecular signatures database.

Figure 5 | Pathway analysis identifies key genes in known pathways which are altered in relation to specific conditions tested in the experiment. Genes can be up- or down-regulated leading to activated or deactivated pathways. Known interactions between pathways can help decipher signaling or regulatory cascades that are affected under the tested condition(s) and regulatory networks can be built towards the understanding of physiological changes caused by the perturbation that was tested.

3. Quantification

Using Sample-to-sample Correlation and Principal Component Analysis to QC Differential Gene Expression

As precision and accuracy of statistical testing is influenced by the reproducibly and variation within the experiment, it is useful to assess the overall similarity between samples when performing differential gene expression testing. Additionally, one should explore if the observed variation is indeed predominantly caused by the experimental conditions or influenced by other technical or biological aspects (e.g., the day of RNA isolation, operator, age, or sex of the organism etc.) This can be investigated by examining sample-to-sample correlations and by performing a principal component analysis (PCA, Fig. 6).

Besides differential gene expression statistics, analysis software such as DESeq2 can also produce normalized and linearized expression data. This output can be used to calculate correlation coefficients between samples. Plotting these values in the form of a clustered heatmap is a quick and visually intuitive way to assess reproducibly between replicates and check for extreme outliners. On the other hand, PCA is a transformation technique which aims to reduce the dimensionality of data while retaining maximal variation in the data set. Gene expression data is highly dimensional, as each sample consists of several thousand data points. During PCA analysis, the information contained in these dimensions is transformed into separate uncorrelated vectors (i.e., principal components). Principle components are ordered in a way in which the first few retain most of the variation present in the original dimensions. Therefore, plotting the two principal components allow the user to obtain a summary of the variation present in the experiment (See here for a more detailed and visual explanation). In this plot (Fig. 6), one can investigate if the assigned samples groups stratify according to the experimental setup (control vs. treatment) or based on other properties (RNA isolation day, age, sex etc.). When sample groups stratify in relation to the other properties listed above, the differentially expressed genes are most likely causally related to them and not the experimental setup. The DESeq2 manual has an excellent tutorial that describes these quality control steps in practice.

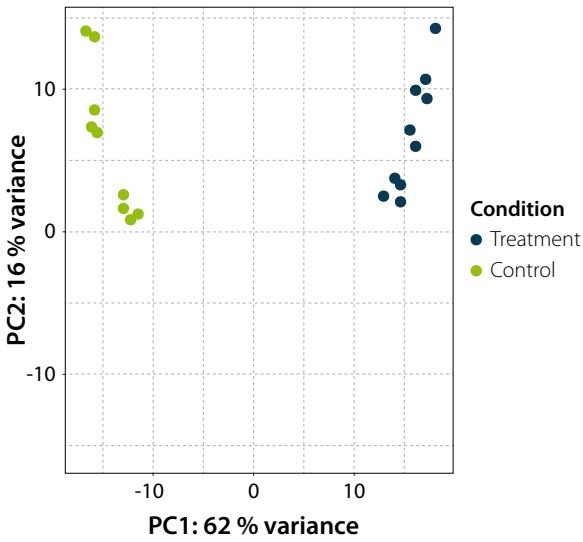


Figure 6 | Principal Component Analysis of treated and untreated control samples. RNA-Seq libraries were produced from two different conditions. Differential expression values were then evaluated in a Principal Component Analysis (PCA). The Principal Component 1 (PC1) on the x-axis clearly separates both conditions, explaining 62 % of the variability seen in the treated vs. untreated control samples clearly separating the two conditions. Replicates from both conditions cluster with less variance in Principle Component 2 (PC2).

Using Spike-in Controls to Validate Sample-to-sample Fold-change

Spike-in controls, such as the ERCCs can be obtained as different mixes. These mixes contain the same spike-ins but at varied concentrations and are helpful when spiked into different sample types (e.g., control vs. treatment). When these samples are then compared during differential gene expression analysis, they can be utilized to quality control fold-change estimates. Similarly, Lexogen's SIRVs, which are available in equimolar or non-equimolar concentrations, can also be used to estimate sample-to-sample fold-changes when spiked in at comparable percentages. In addition, SIRVs contain various synthetic isoforms and thus simulate the full transcriptomic complexity. This feature makes them very useful to test RNA-Seq and data analysis workflows especially for evaluations on transcript level.

4. Tying it all together

Combining the aforementioned analysis steps into a single automated workflow is referred to as a "pipeline". In the past, bioinformaticians mainly wrote pipelines tailored to their specific systems and needs. This was most commonly done using the UNIX shell programming language, bash. However, as the field of bioinformatics has rapidly matured, the demand for reproducible and sharable data analysis workflows has greatly increased. This demand has led to the development of sophisticated pipeline managers such as Snakemake or Nextflow, which enable greater reproducibility and ease of sharing.

NextFlow is also utilized by the nf-core project[10] which is a community effort to build curated data analysis pipelines for various NGS sequencing applications. These pipelines are open source, well tested, and adhere to stringent quality standards. They provide an excellent starting point for researchers new to NGS data analysis and can be downloaded via the nf-core page. As this is a community effort, it is highly encouraged that researchers contribute and share their own pipelines!

5. Some Actionable Advice

After carefully analyzing your data and controlling the individual step to match commonly defined statistics and expected results (e.g., for spike-in controls), researchers are well equipped to proceed to visualizing their data and providing conclusive arguments to solve their individual research question.

To ensure the success of a sequencing project early on, it is highly recommended to consult with an experienced bioinformatician already during the experimental planning stages or revert to a service provider to discuss the project. For researchers without bioinformatics staff or experience in data analysis, third party data analysis platforms provide a convenient solution and allow researchers to analyze their own data using validated pipelines.

Lexogen also offers RNA-Seq data analysis service and provides intuitive plug-and-play data analysis pipelines on our partner platforms. For more information and a comprehensive overview of the various pipelines visit our FAQ page on Data Analysis Solutions and consult with us!

Literature:

1. Boothby, T. C., Tenlen, J. R., Smith, F. W., *et al.*, (2015) Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc. Natl. Acad. Sci.* 112: 15976 15981. doi:10.1073/pnas.1510461112

2. Bemm, F., Weiß, C. L., Schultz, J., Förster, F. (2016) Genome of a tardigrade: Horizontal gene transfer or bacterial contamination? *Proc. Natl. Acad. Sci.* 113: E3054-E3056. doi:10.1073/pnas.1525116113

3. Wang, L., Wang, S., and Li, W., (2012) RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 28: 2184–2185. doi:10.1093/bioinformatics/bts356

4. Zhao S, Zhang Y, Gamini R., et al., (2010) Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci Rep.* 8:4781. doi:10.1038/s41598-018-23226-4

5. Quast, C., Pruesse, E., Yilmaz, P., et al., (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucl. Acids Res.* 41 (D1): D590-D596. doi:10.1093/nar/gks1219

6. The External RNA Controls Consortium (2005) The External RNA Controls Consortium: a progress report. *Nat Methods* 2: 731–734. doi:10.1038/nmeth1005-731

7. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15: 550. doi:10.1186/s13059-014-0550-8

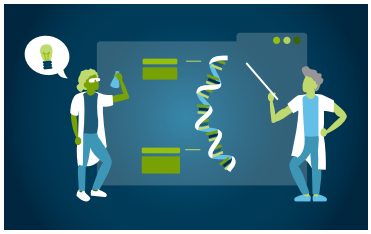
8. Robinson, M. D., McCarthy, D. J., and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139-140. doi:10.1093/bioinformatics/btp616

9. Yu, G., Wang, L.-G., Yanyan Han, Y., and He, Q-Y. (2012) clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS: *A Journal of Integrative Biology.* 16:284-287. doi:10.1089/omi.2011.0118

10. Ewels, P.A., Peltzer, A., Fillinger, S. et al. (2020) The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol* 38:276–278. doi:10.1038/s41587-020-0439-x

LEXOGEN
The RNA Experts

Curious to learn more?

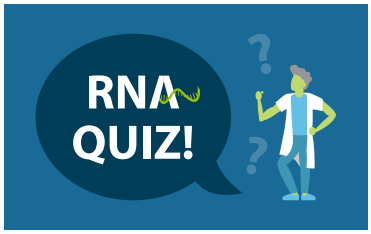


Explore more chapters in our RNA LEXICON:

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 Quizzes:

- Quiz 1 (Chapters # 1 - 3): www.lexogen.com/lexicon-quiz-1
- Quiz 2 (Chapters # 4 - 6): www.lexogen.com/lexicon-quiz-2
- Quiz 3 (Chapters # 7 - 8): www.lexogen.com/lexicon-quiz-3
- Quiz 4 (Chapters # 9 - 10): www.lexogen.com/lexicon-quiz-4
- Quiz 5 (Chapters # 11 - 13): www.lexogen.com/lexicon-quiz-5

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