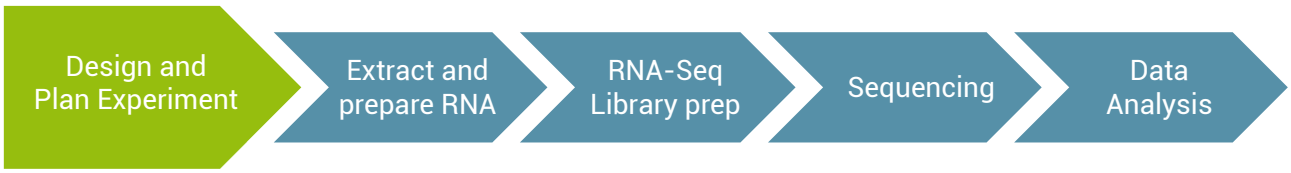
LEXOGEN
The RNA Experts

RNA LEXICON

CHAPTER #3

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.

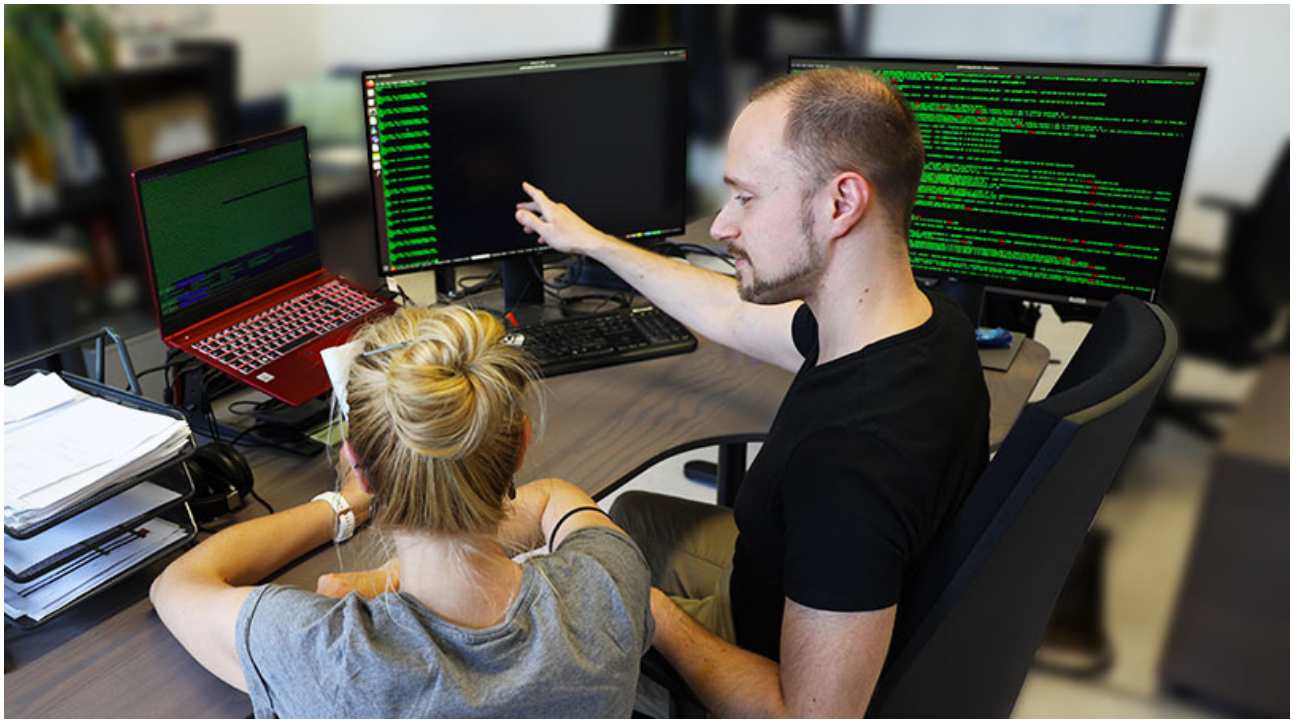
The choice of library preparation impacts the kind of data that can be obtained in the RNA-Seq experiment and ultimately needs to be aligned with the level of information that is required. Apart from sample type-related constrains, e.g., poly(A) selection is not possible for organisms that do not contain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

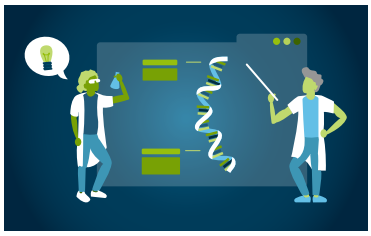
After carefully planning the experiment, it is time to prepare the samples for sequencing. Before starting a large-scale or long-term experiment, it is extremely useful to conduct a pilot experiment with a representative but smaller set of samples to check if the chosen experimental parameters deliver 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.

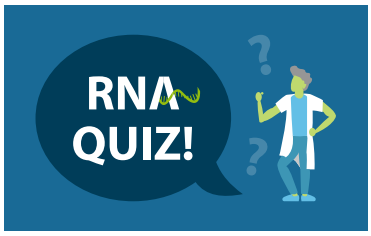
LEXOGEN
The RNA Experts

Curious to learn more?



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

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



Show your RNA expertise and master all questions of our RNA Quiz:
www.lexogen.com/lexicon-quiz-1

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