NGS library preparation

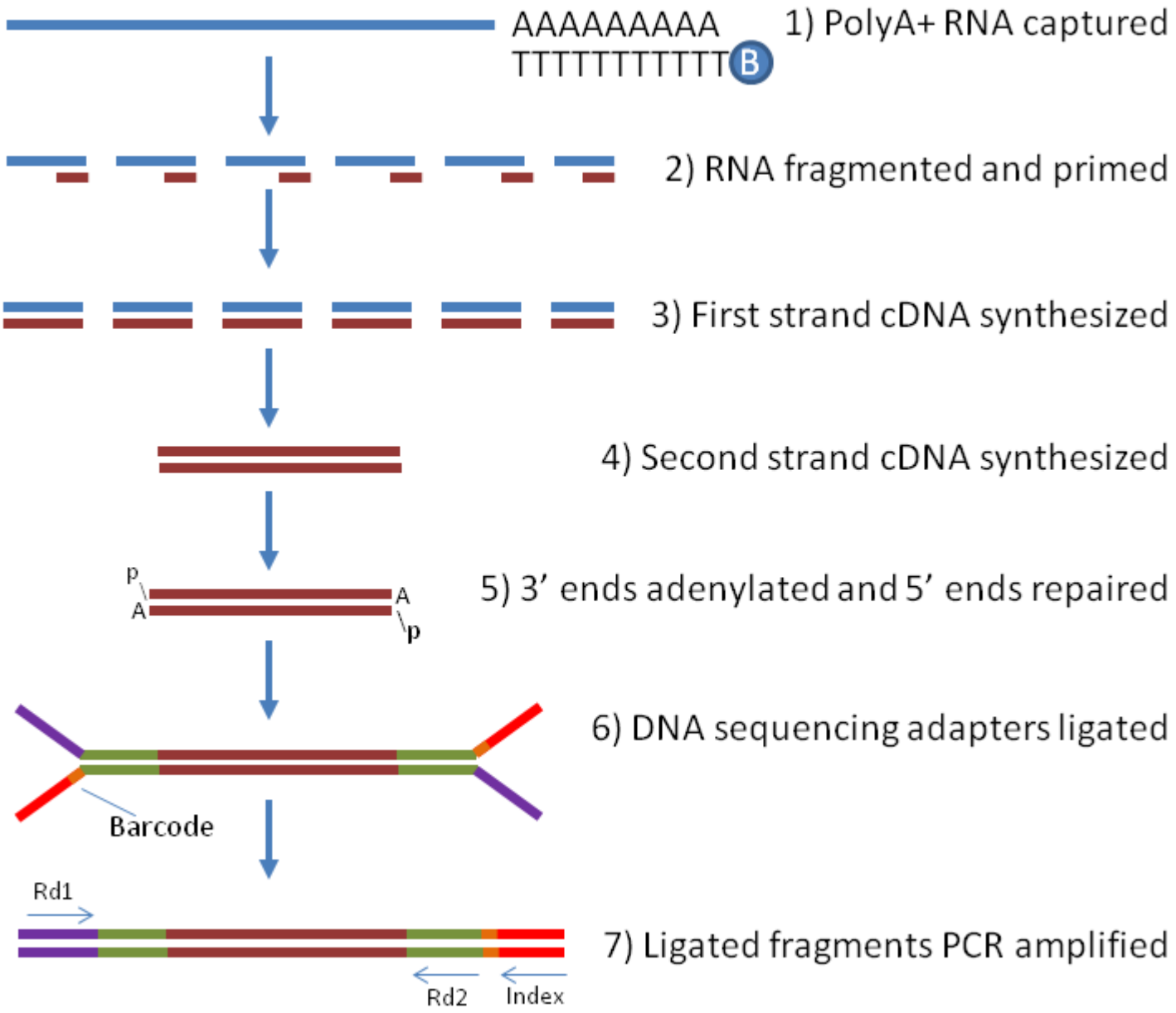
RNA

- The regulation of RNA transcription and processing directly affects protein synthesis. Proteins, in turn, mediate cellular functions to establish the phenotype of the cell. Dysregulated RNAs are the cause for some diseases and cancers.
- Sequencing RNA provides information about both the abundance and sequence of the RNA molecules.

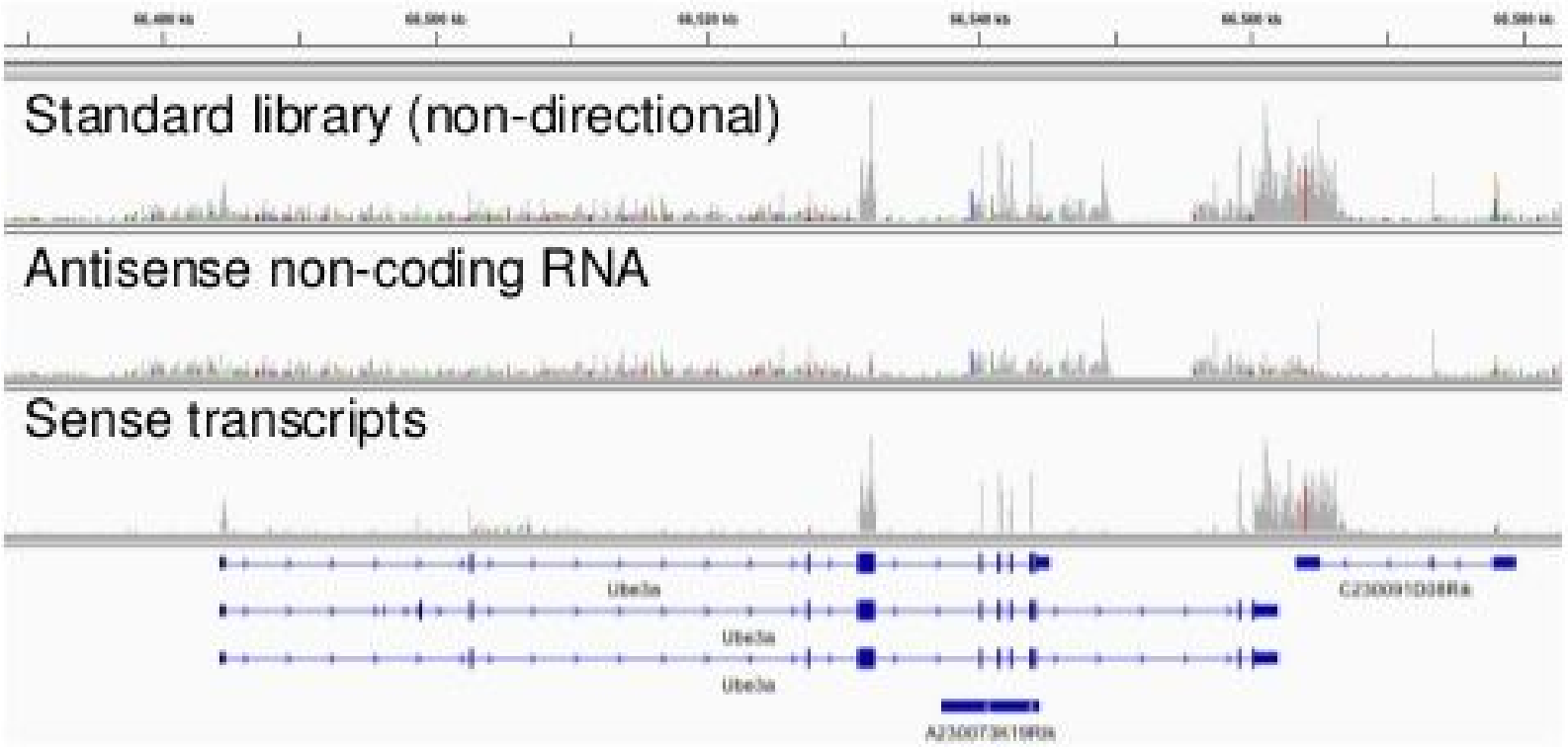
RNA-seq produces millions of sequences from complex RNA samples. With this powerful approach, you can:

1. Measure gene expression.
2. Discover and annotate complete transcripts.
3. Characterize alternative splicing and polyadenylation.

RNA-SEQ

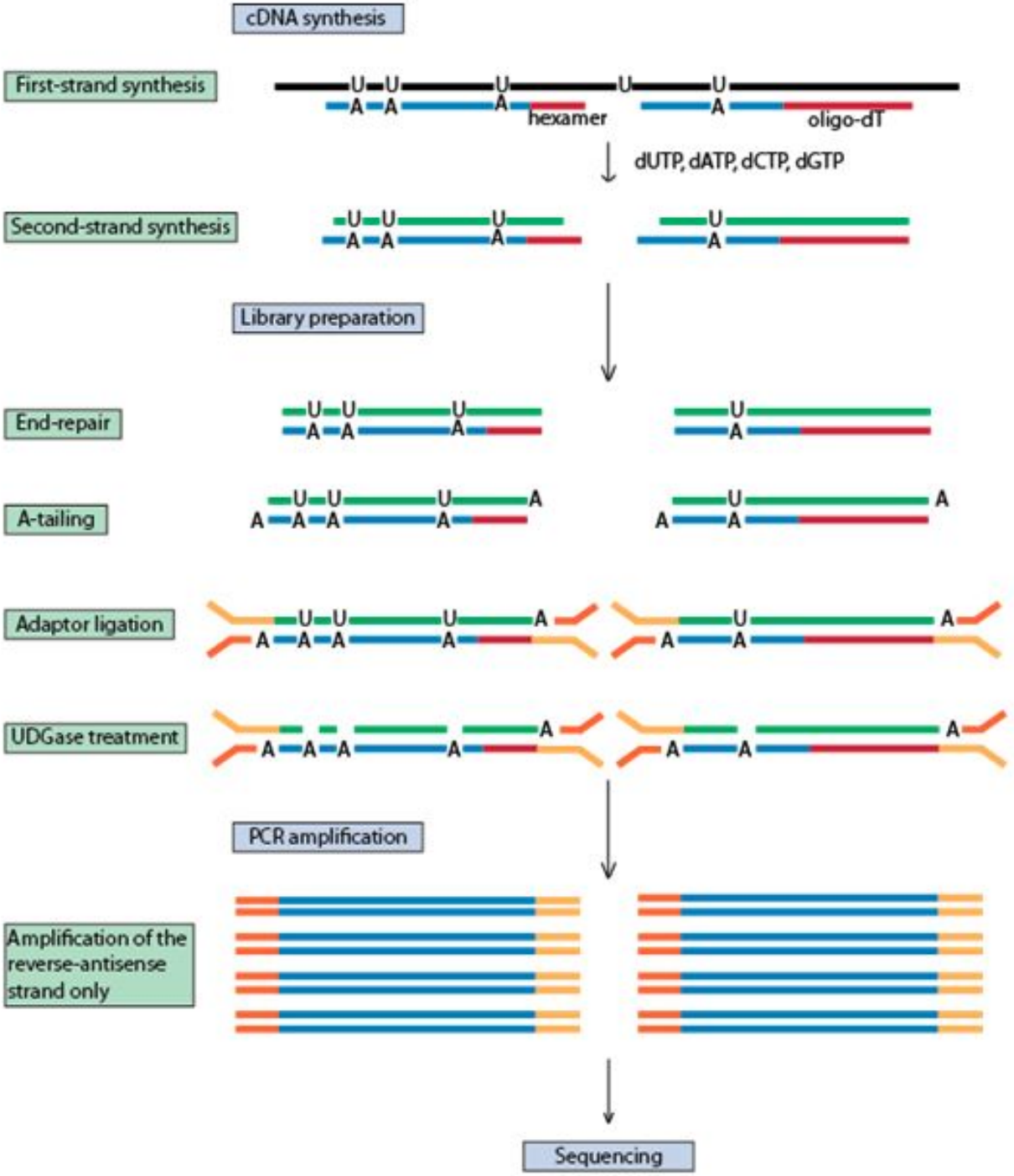

AAAAAAAAA
TTTTTTTTTTB
1) PolyA+ RNA captured
2) RNA fragmented and primed
3) First strand cDNA synthesized
4) Second strand cDNA synthesized
p
A
A
p
5) 3’ ends adenylated and 5’ ends repaired
6) DNA sequencing adapters ligated
Barcode
Rd1
Rd2
Index
7) Ligated fragments PCR amplified

Strand-specific RNA-seq

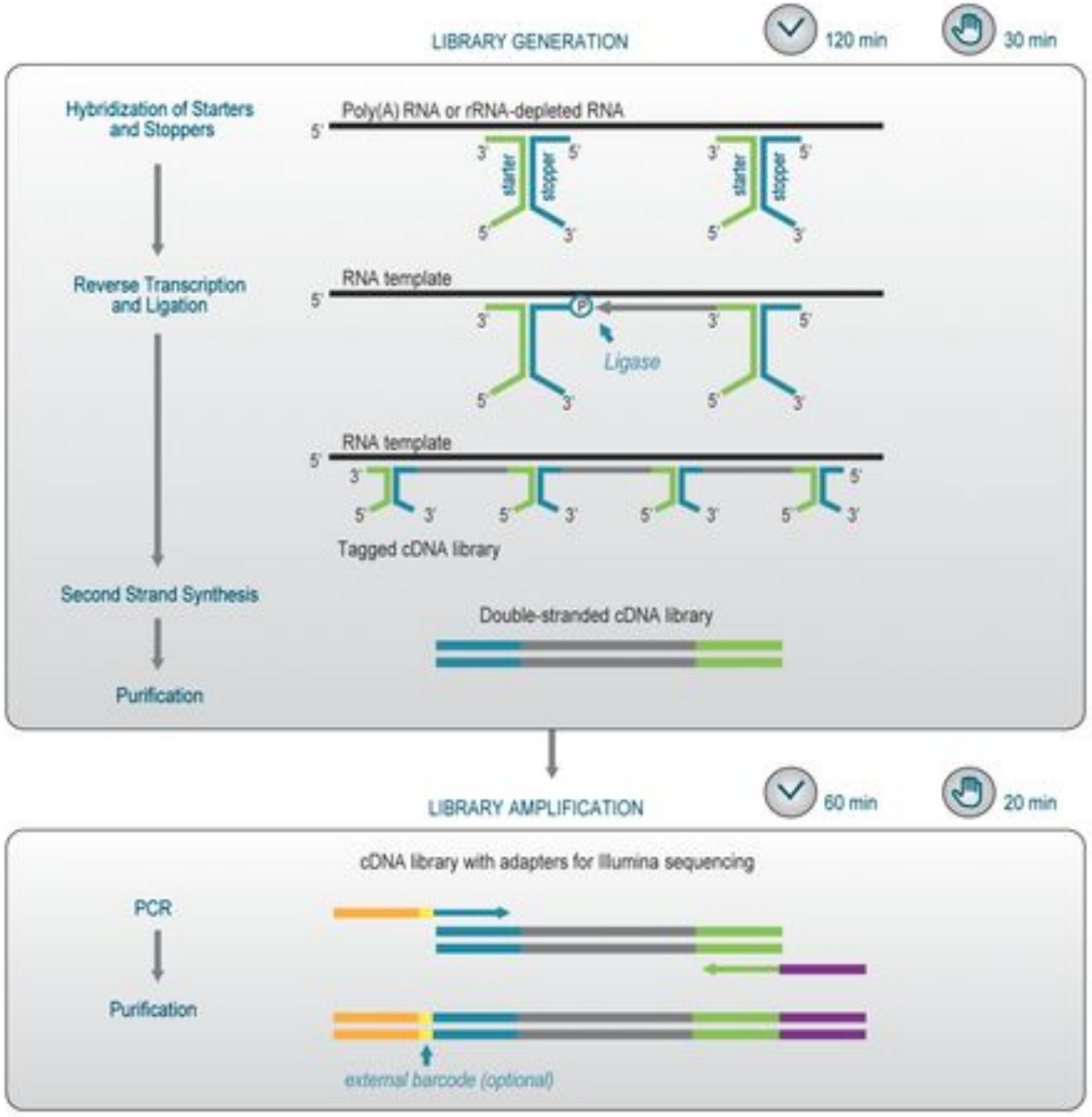


- Informative for non-coding RNAs and antisense transcripts
- Essential when NOT using polyA selection (mRNA)
- No disadvantage to preserving strand specificity

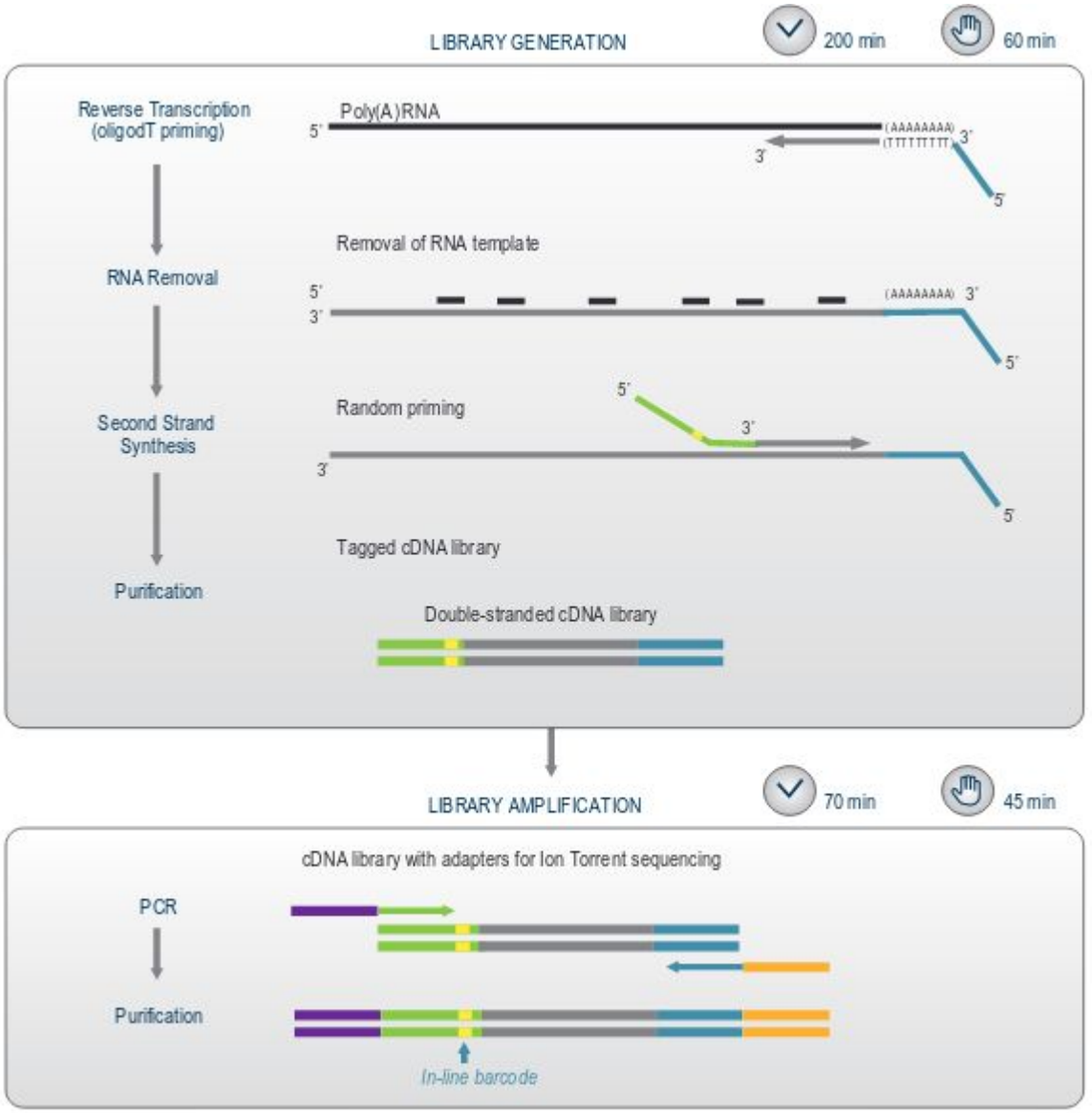
Librerie unidirezionali



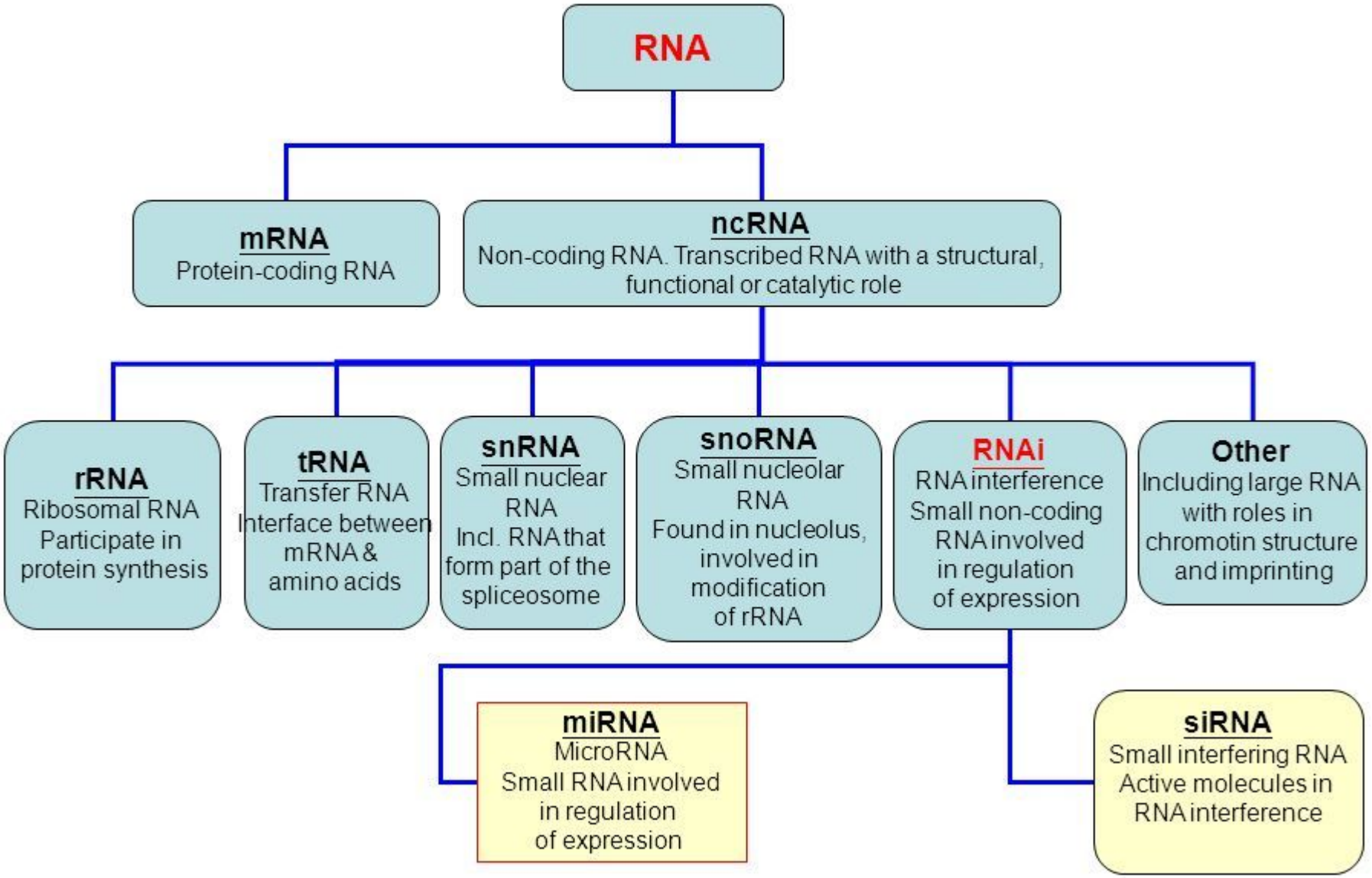
Metodi alternativi



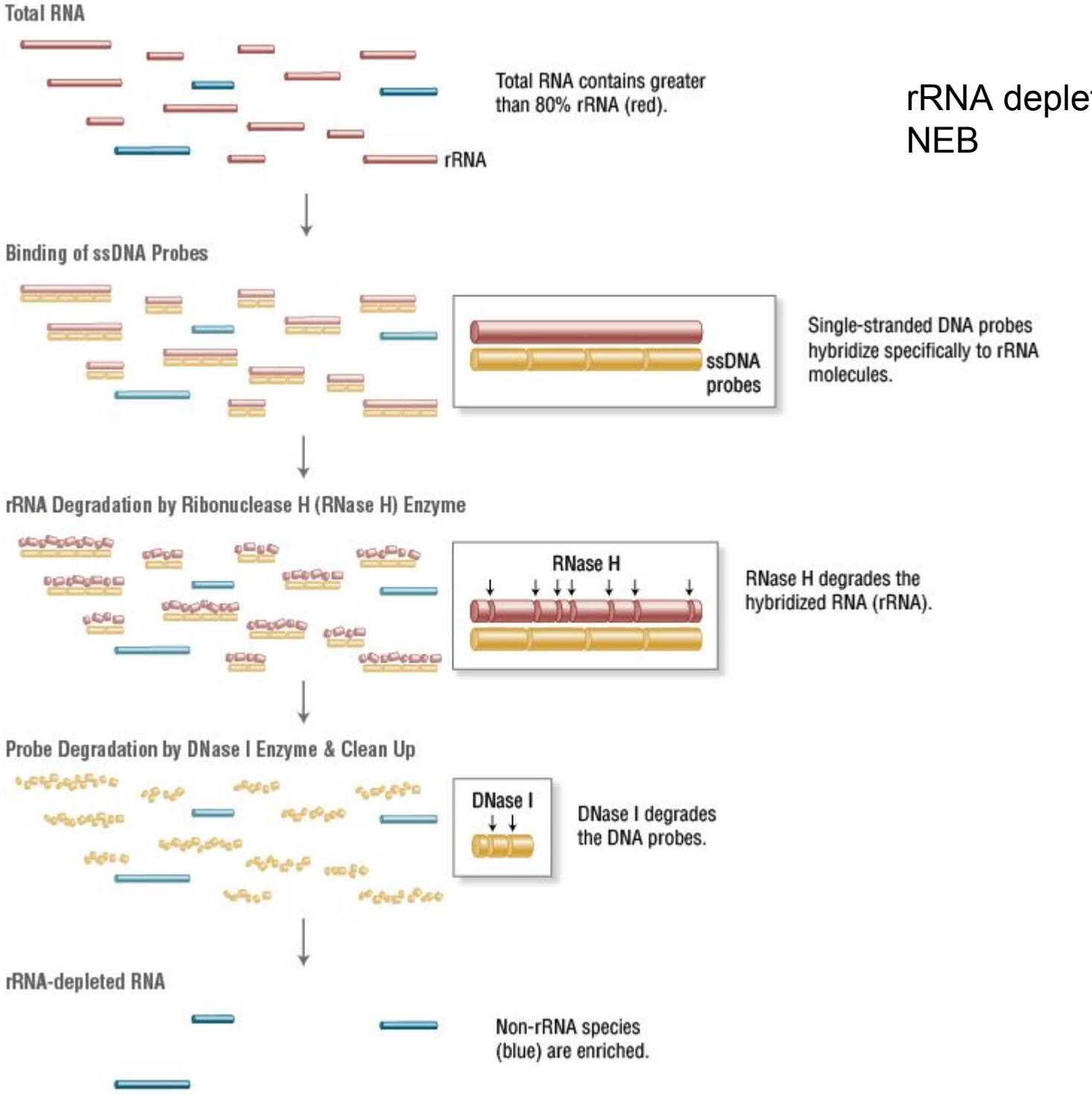
3’ seq

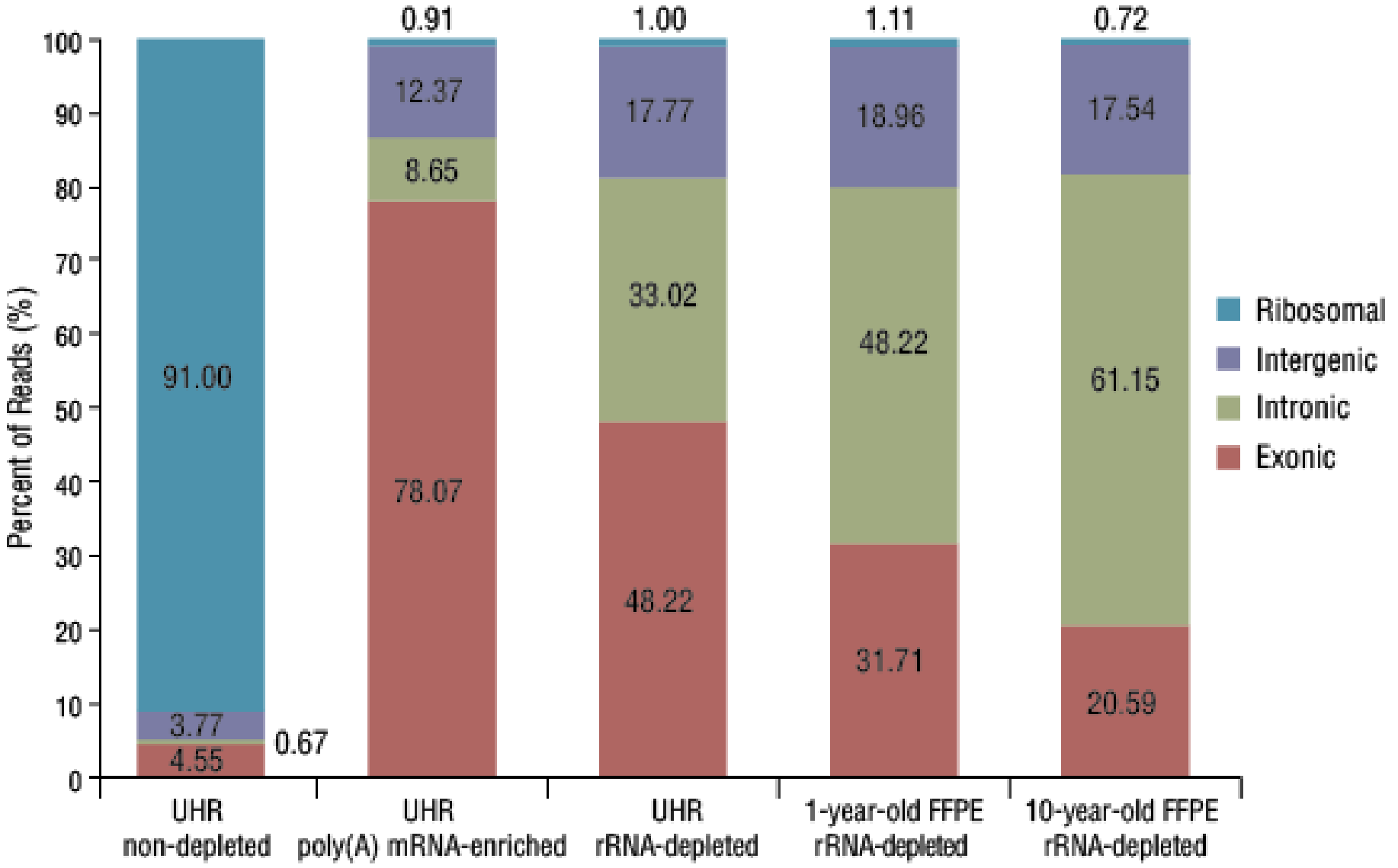

LIBRARY GENERATION
200 min
60 min
Reverse Transcription
(oligodT priming)
Poly(A)RNA
RNA Removal
Removal of RNA template
Second Strand
Synthesis
Random priming
Purification
Tagged cDNA library
Double-stranded cDNA library
LIBRARY AMPLIFICATION
70 min
45 min
cDNA library with adapters for Ion Torrent sequencing
PCR
Purification
In-line barcode

Type of RNA molecules

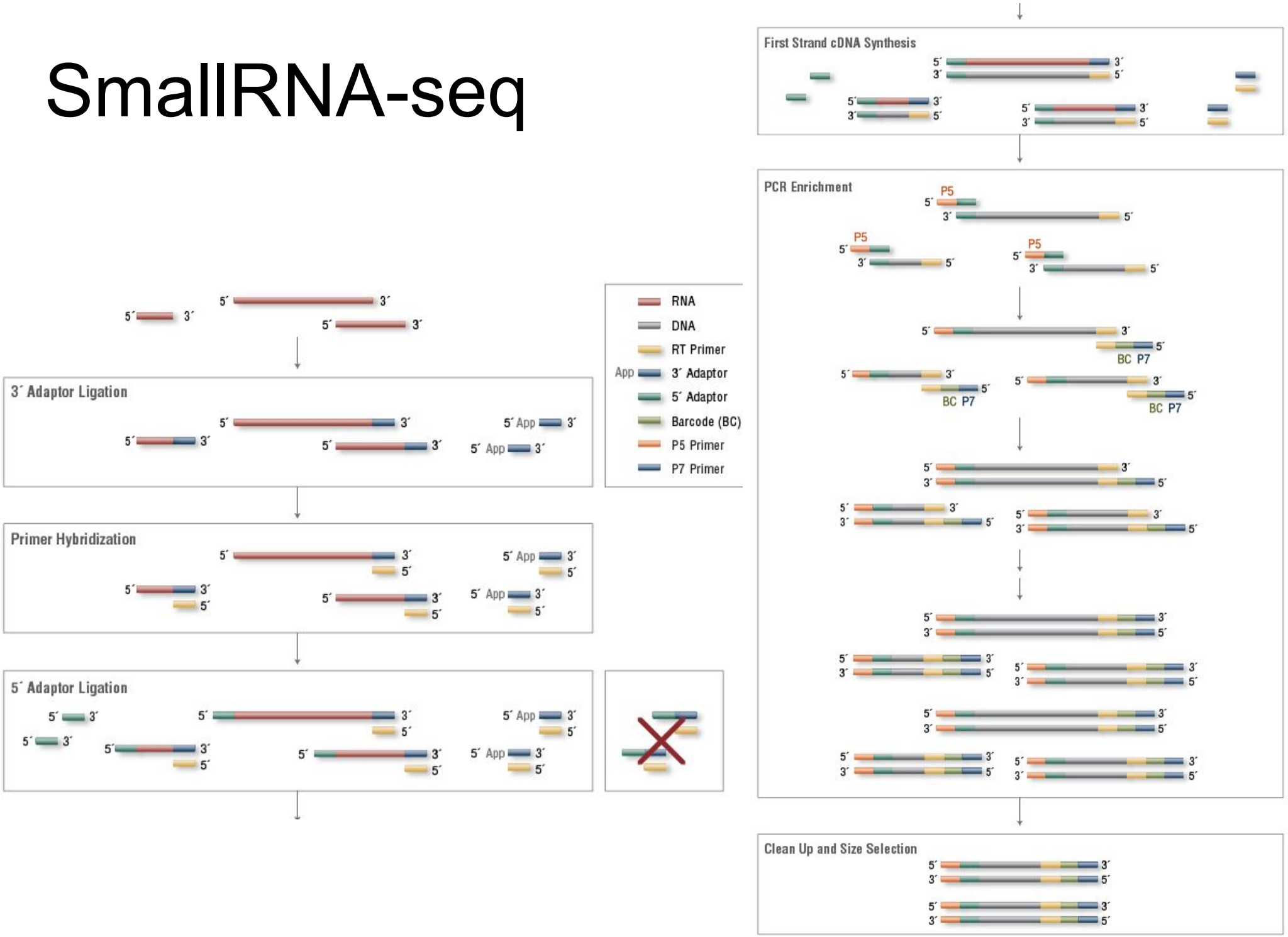


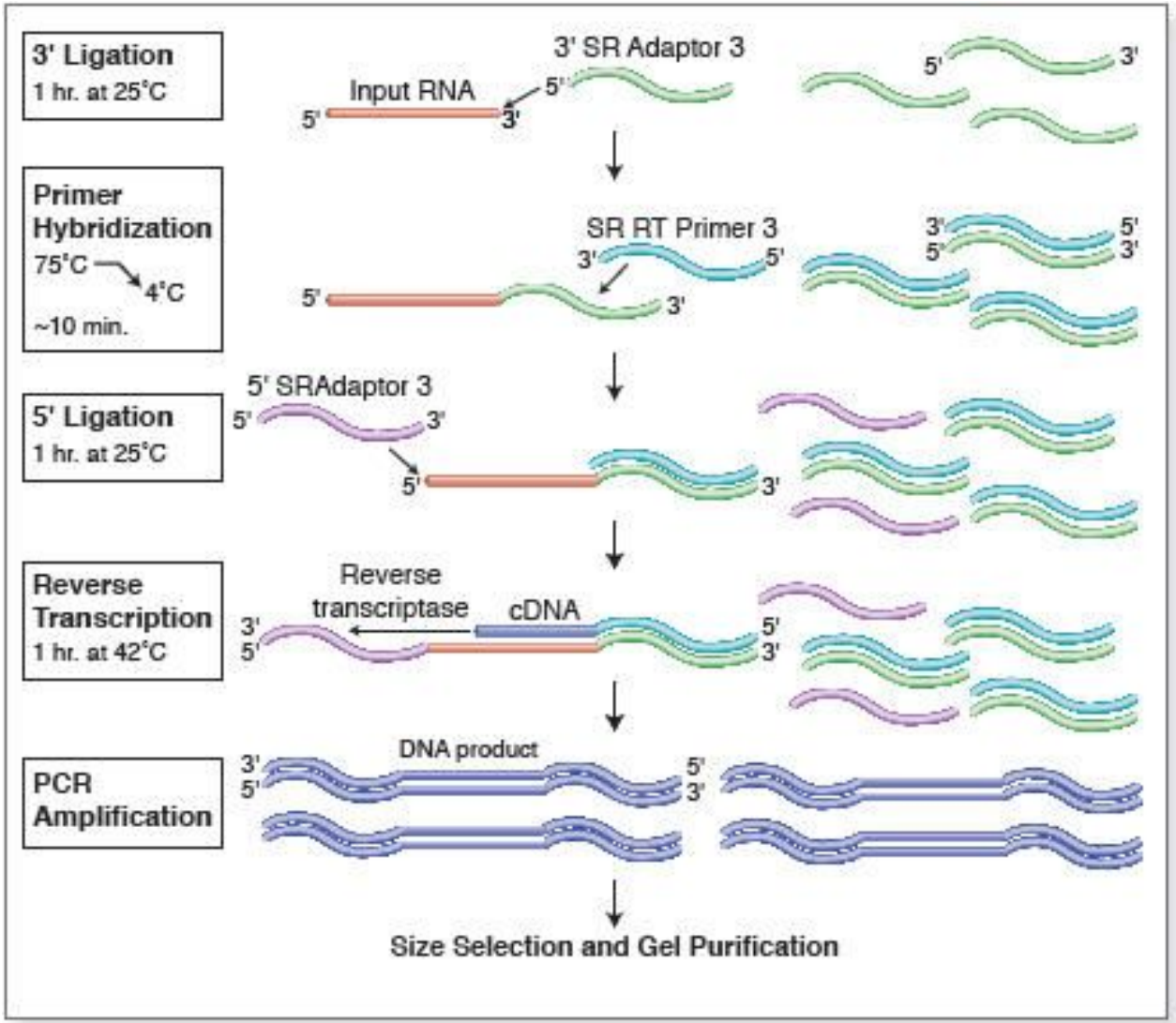
rRNA depletion
NEB

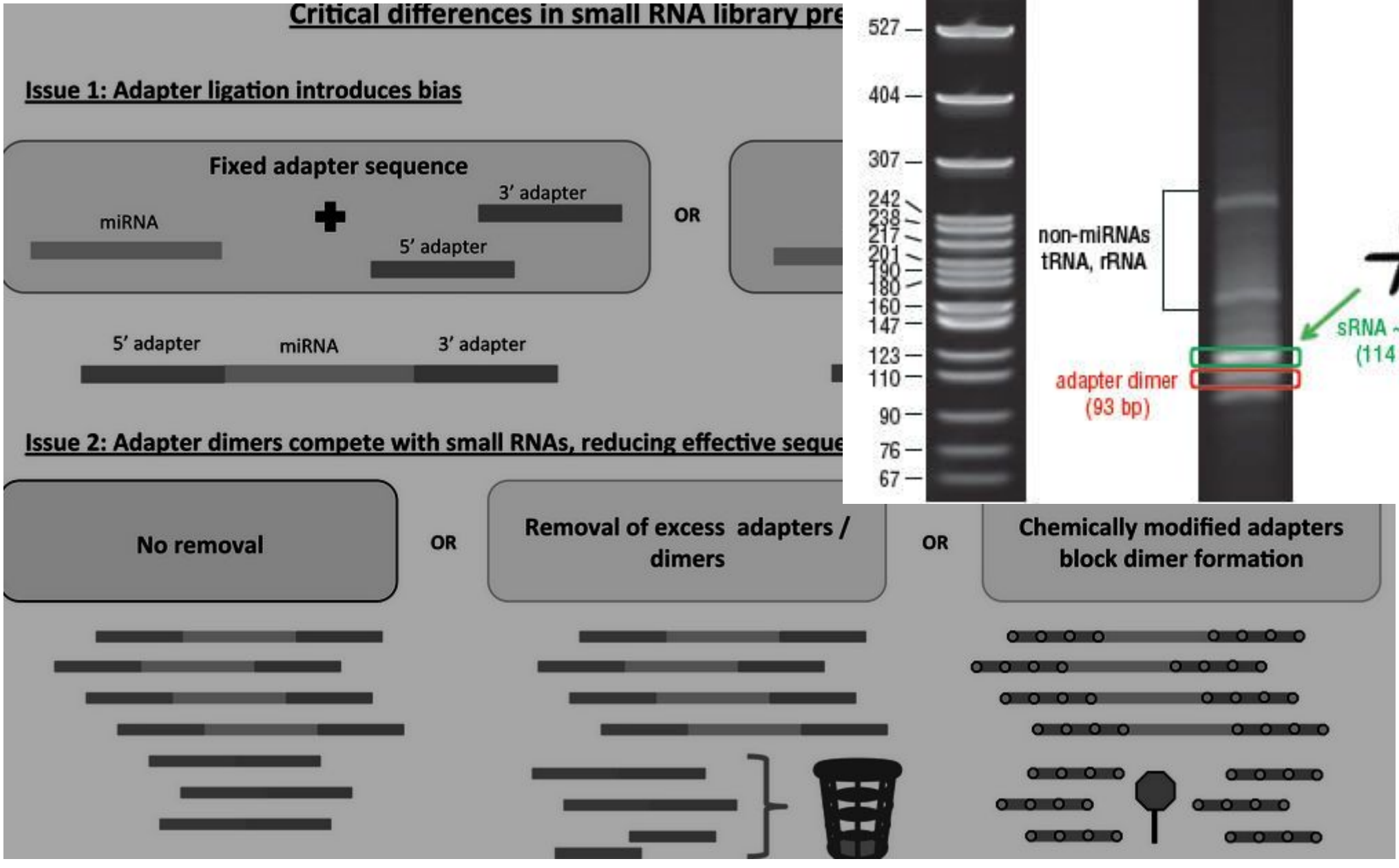



Percent of Reads (%)
100
90
80
70
60
50
40
30
20
10
0
91.00
3.77
4.55
0.67
0.91
12.37
8.65
78.07
1.00
17.77
33.02
48.22
48.22
1.11
18.96
48.22
31.71
0.72
17.54
61.15
20.59
Ribosomal
Intergenic
Intronic
Exonic
UHR non-depleted
UHR poly(A) mRNA-enriched
UHR rRNA-depleted
1-year-old FFPE rRNA-depleted
10-year-old FFPE rRNA-depleted

SmallRNA-seq




3' Ligation
1 hr. at 25°C
3' SR Adaptor 3
Input RNA
5'
3'
Primer Hybridization
75°C
4°C
~10 min.
SR RT Primer 3
5' SRAdaptor 3
5' Ligation
1 hr. at 25°C
Reverse transcriptase
cDNA
Reverse Transcription
1 hr. at 42°C
DNA product
PCR Amplification
Size Selection and Gel Purification


Critical differences in small RNA library pre
Issue 1: Adapter ligation introduces bias
Fixed adapter sequence
miRNA
3' adapter
5' adapter
OR
5' adapter
miRNA
3' adapter
Issue 2: Adapter dimers compete with small RNAs, reducing effective seque
No removal
OR
Removal of excess adapters / dimers
OR
Chemically modified adapters block dimer formation

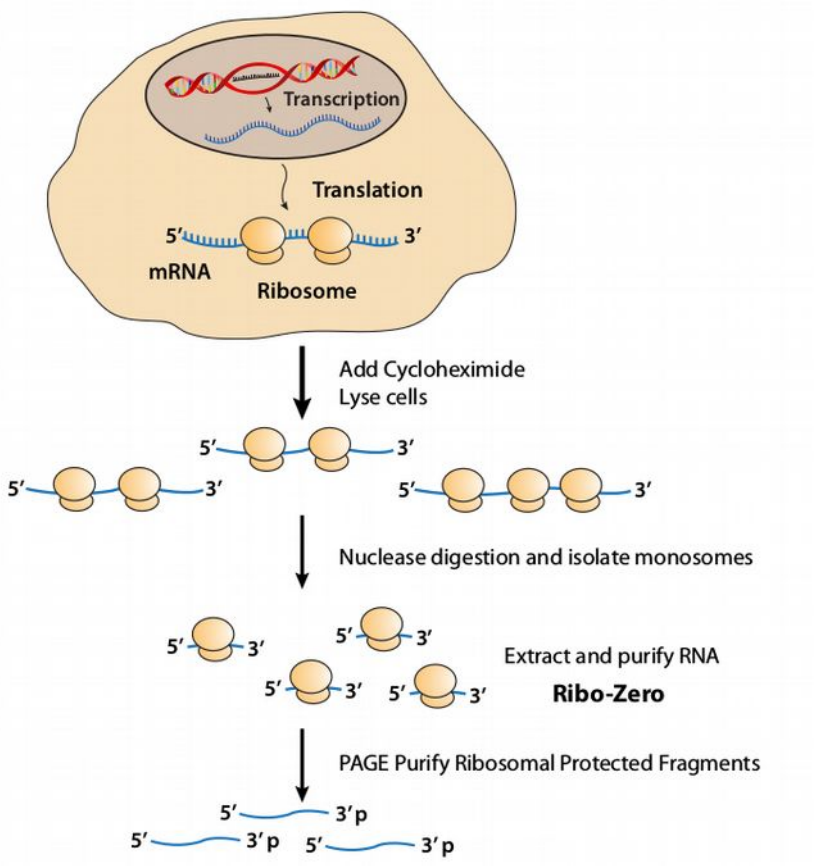
Total RNA
Lane 1
Quick-Load
pBR322
DNA-MspI Digest
Lane 2
Human
Brain
bp
622
527
404
307
242
238
217
201
190
180
160
147
123
110
90
76
67
non-miRNAs
tRNA, rRNA
sRNA ~21 nt
(114 bp)
adapter dimer
(93 bp)

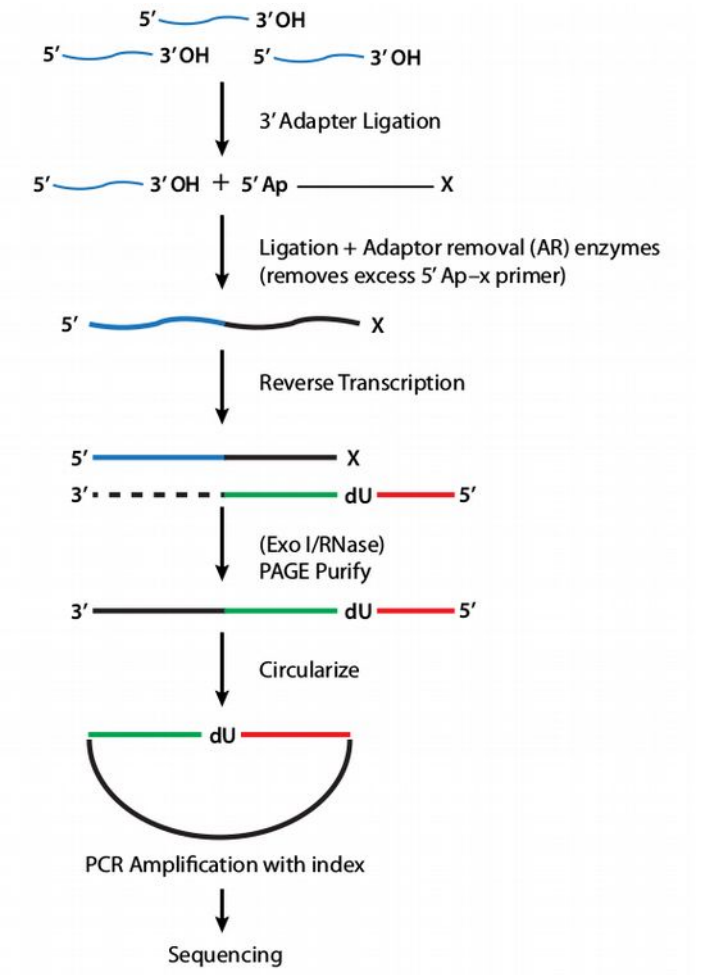
Not only gene expression

- RNA modification,
- RNA-binding protein domain
- RNA structure

RIBOSOME PROFILING SEQUENCING (RIBO-SEQ)/ARTSEQ

- Active mRNA Translation Sequencing (ART-seq), also called ribosome profiling (Ribo-Seq), isolates RNA that is being processed by the ribosome in order to monitor the translation process.
- The addition of harringtonine (an alkaloid that inhibits protein biosynthesis) causes ribosomes to accumulate precisely at initiation codons and assists in their detection.





Ribosome profiling more closely reflects the rate of protein synthesis than mRNA levels

RNA IMMUNOPRECIPITATION SEQUENCING (RIP-SEQ)

- RNA immunoprecipitation sequencing (RIP-Seq) maps the sites where proteins are bound to the RNA within RNA-protein complexes.
- In this method, RNA-protein complexes are immunoprecipitated with antibodies targeted to the protein of interest.

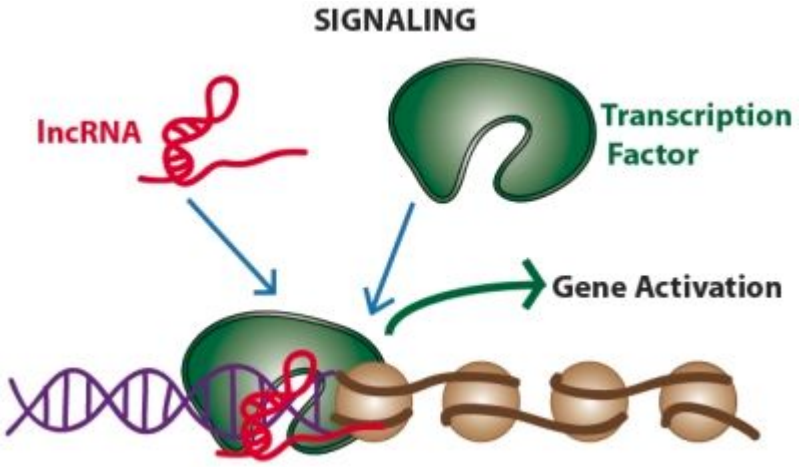
- After RNase digestion, RNA covered by protein is extracted and reverse-transcribed to cDNA. The locations can then be mapped back to the genome. Deep sequencing of cDNA provides single-base resolution of bound RNA.

Maps specific protein-RNA complexes, such as polycomb-associated RNAs

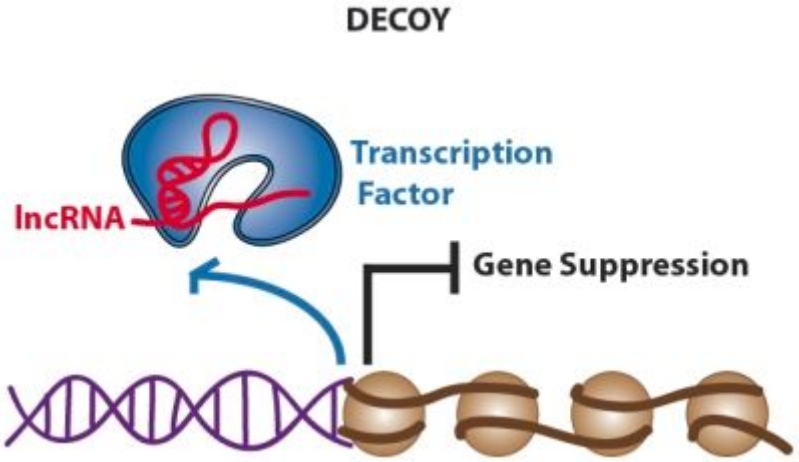
HIGH-THROUGHPUT SEQUENCING OF CLIP CDNA LIBRARY (HITS-CLIP)
OR
CROSSLINKING AND IMMUNOPRECIPITATION SEQUENCING (CLIP-SEQ)

- This approach is similar to RIP-Seq, but uses crosslinking to stabilize the protein-RNA complexes. In this method, RNA-protein complexes are UV crosslinked and immunoprecipitated. The protein-RNA complexes are treated with RNase followed by Proteinase K.

Long Non-Coding RNAs (lncRNAs) Functional Analysis

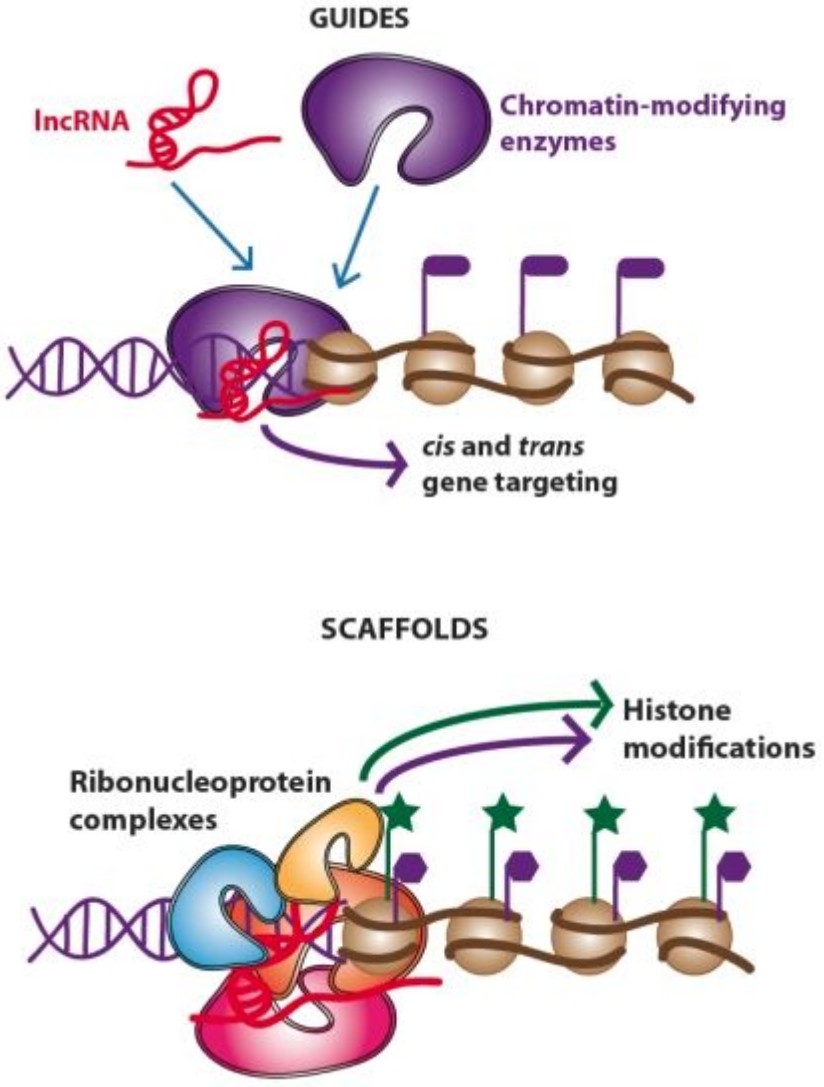


- Long non-coding RNAs acting as gene activators (signaling archetype).



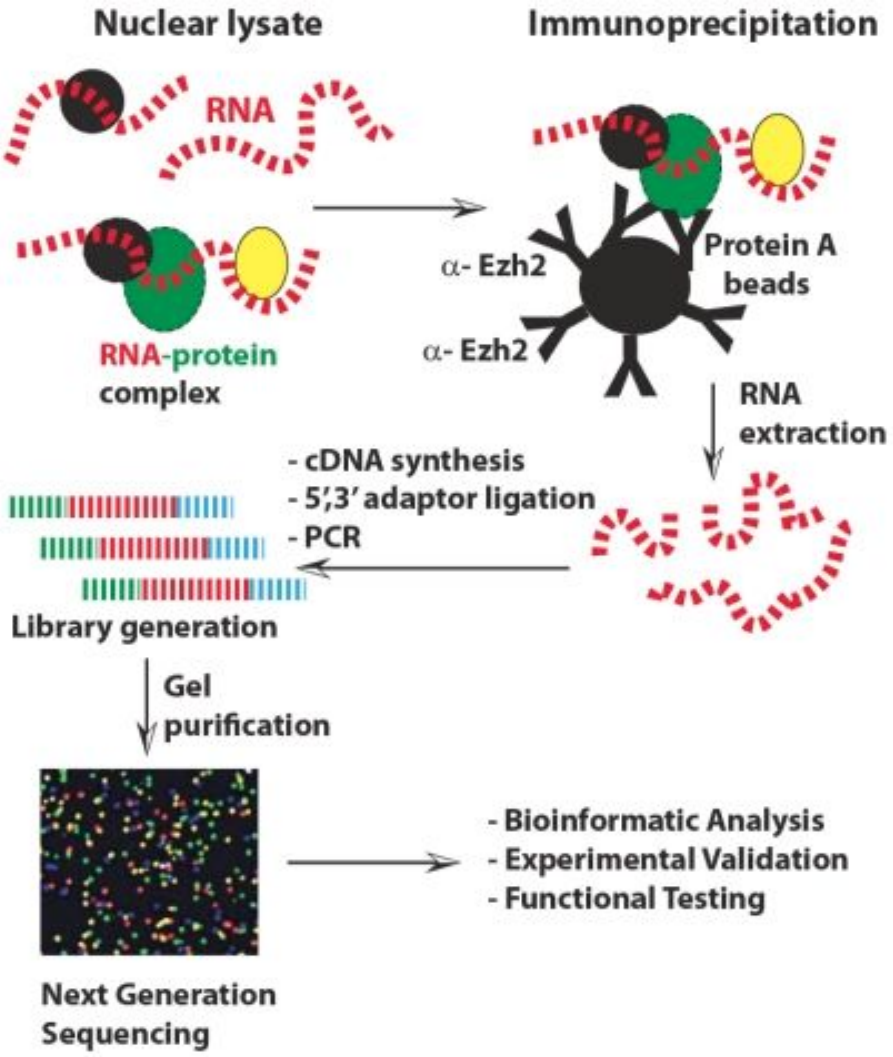
- Long non-coding RNAs acting as gene suppressors (decoy archetype).

Long Non-Coding RNAs (lncRNAs) Functional Analysis



- Long non-coding RNAs acting as cis and trans gene expression regulators (guide archetype).
-
- Long non-coding RNAs acting as chromatin modificators (scaffold archetype).

Long Non-Coding RNAs (lncRNAs) Functional Analysis

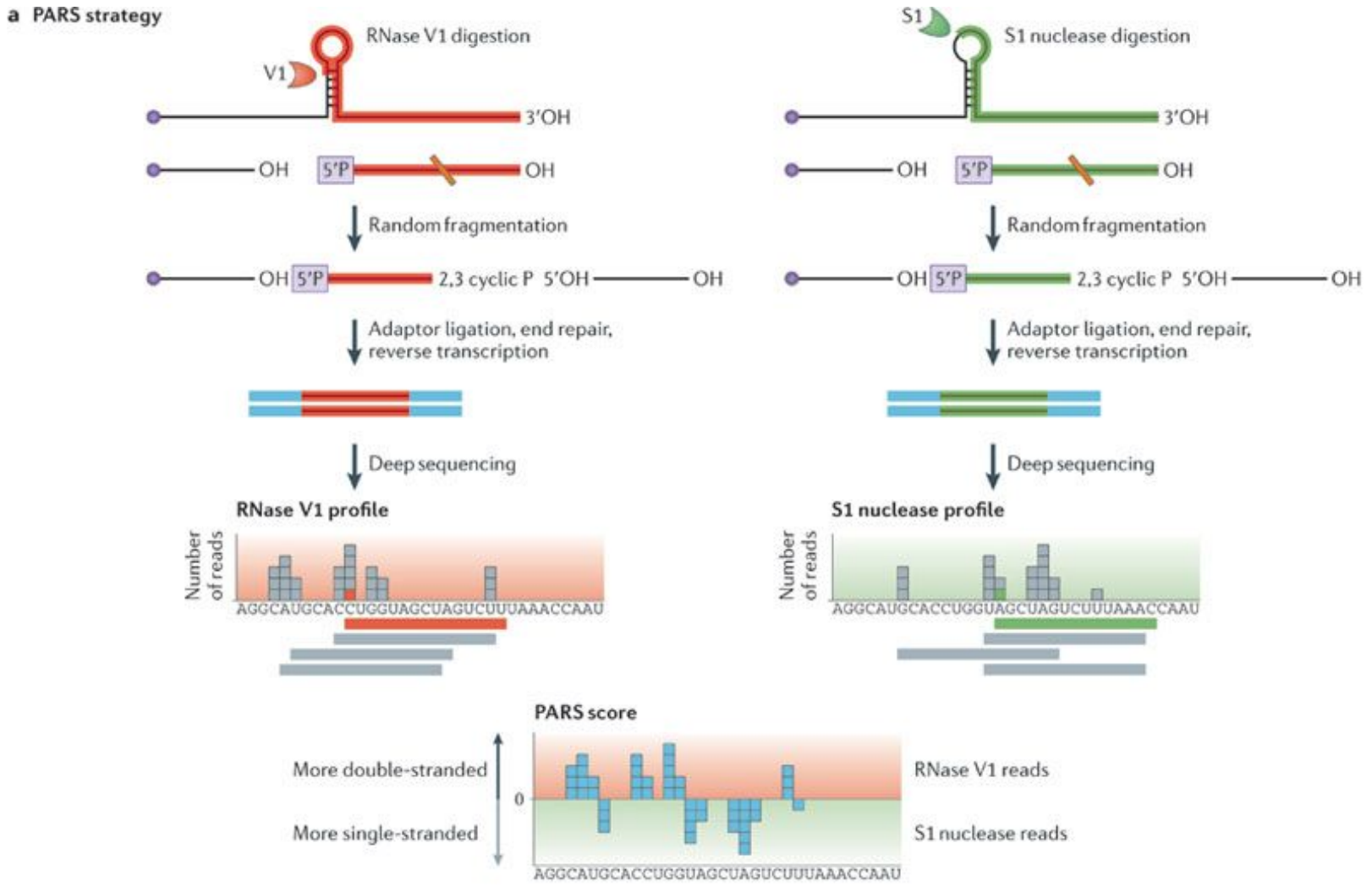


- After cell lysis an immunoprecipitation step is required to isolate RNA bound to Polycomb proteins. Retrotranscription to cDNA and ligation with suitable adapters are required prior to next generation sequencing. Bioinformatics analysis and further validation led to the identification of novel lncRNAs.

RNA STRUCTURE

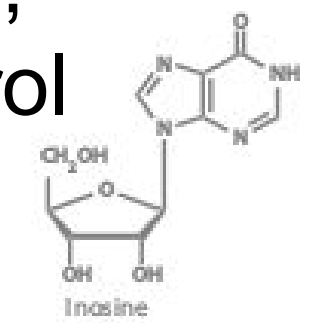
- RNA has the ability to form secondary structures that can either promote or inhibit RNA-protein or protein-protein interactions.
- The most diverse secondary and tertiary structures are found in transfer RNAs (tRNAs) and are thought to play a major role in modulating protein translation.

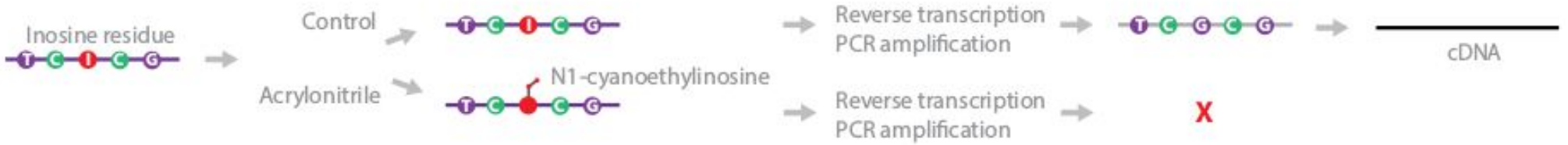
PARALLEL ANALYSIS OF RNA STRUCTURE (PARS-SEQ)



INOSINE CHEMICAL ERASING SEQUENCING (ICE)

- Inosine chemical erasing (ICE) identifies adenosine to inosine editing. In this method, RNA is treated with acrylonitrile, while control RNA is untreated.
- Inosines in RNA fragments treated with acrylonitrile cannot be reverse-transcribed.

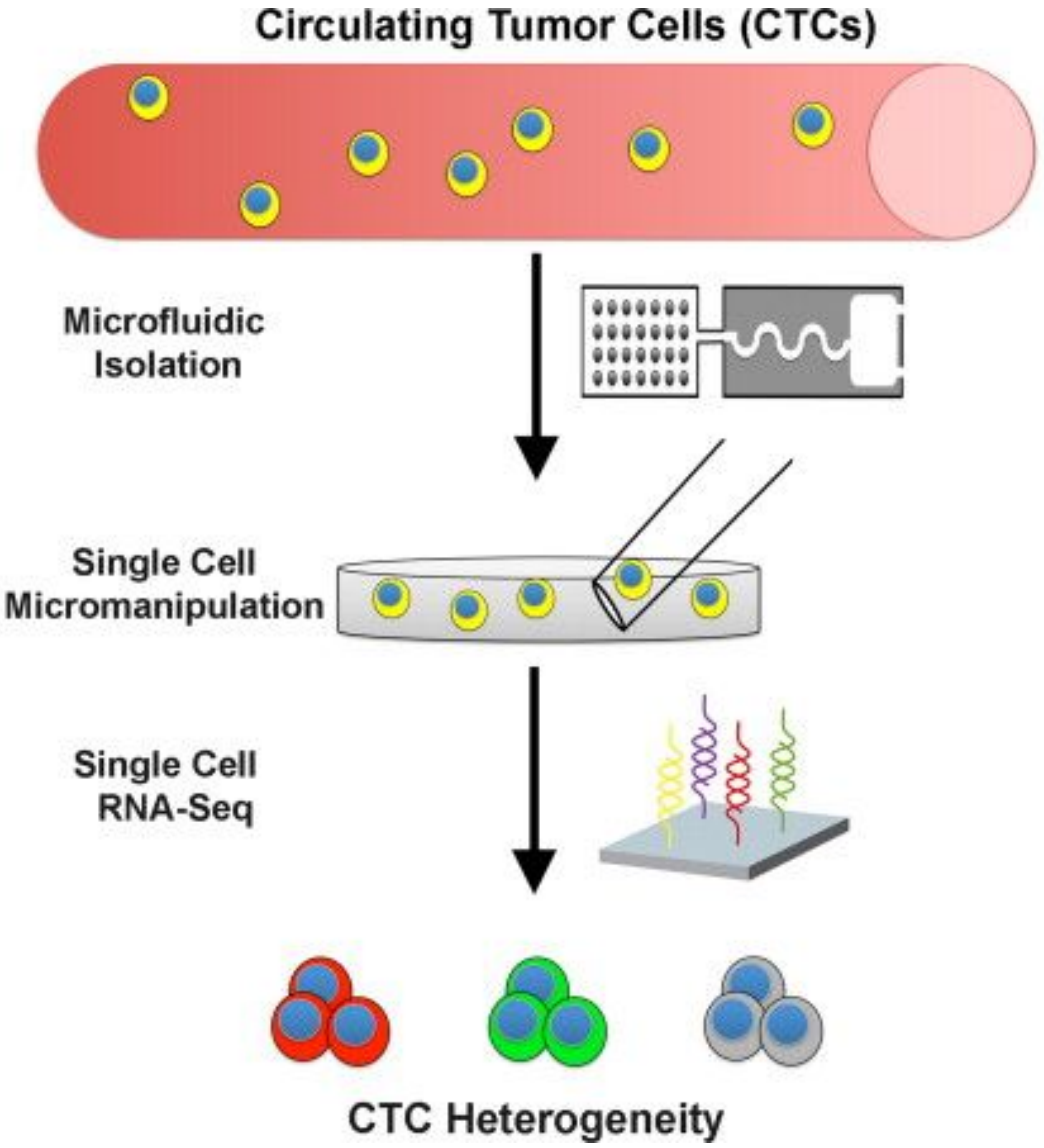




LOW-LEVEL RNA DETECTION

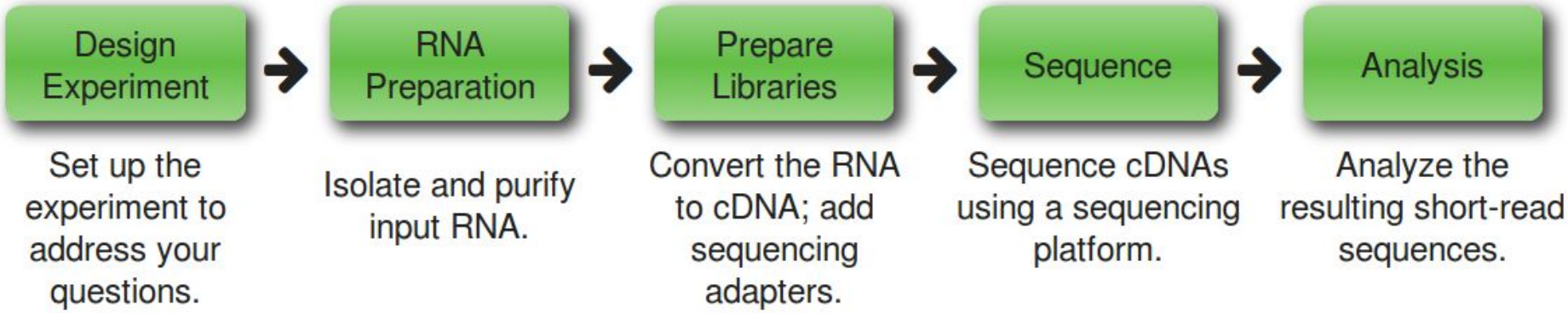
- Low-level RNA detection refers to both detection of rare RNA molecules in a cell-free environment, such as circulating tumor RNA, or the expression patterns of single cells.
- Tissues consist of a multitude of different cell types, each with a distinctly different set of functions. Even within a single cell type, the transcriptomes are highly dynamic and reflect temporal, spatial, and cell cycle–dependent changes.

-
- The RNA content and RNA make up of a cell depend very much on its developmental stage and the type of cell. To estimate the approximate yield of RNA that can be expected from your starting material, we usually calculate that a typical mammalian cell contains 10–30 pg total RNA.
-
- Approximately 360,000 mRNA molecules are present in a single mammalian cell, made up of approximately 12,000 different transcripts with a typical length of around 2 kb.
- Some mRNAs comprise 3% of the mRNA pool whereas others account for less than 0.1%. These rare or low-abundance mRNAs may have a copy number of only 5–15 molecules per cell.


Circulating Tumor Cells (CTCs)
Microfluidic Isolation
Single Cell Micromanipulation
Single Cell RNA-Seq
CTC Heterogeneity

SWITCH MECHANISM AT THE 5’ END OF RNA TEMPLATES VERSION 2 (SMART-SEQ2)

A typical RNA-seq experiments



Experimental design

- When designing an RNA-seq experiment researchers are faced with choosing between many experimental options, and decisions must be made at each step of the process.

Identify the primary experimental objective

- The types of information that can be gained from RNA-seq can be divided into two broad categories: qualitative and quantitative.
- **Qualitative data** includes identifying expressed transcripts, and identifying exon/intron boundaries, transcriptional start sites (TSS), and poly-A sites. Here, we will refer to this type of information as "annotation".
- **Quantitative data** includes measuring differences in expression, alternative splicing, alternative TSS, and alternative polyadenylation between two or more treatments or groups. Here we focus specifically on experiments to measure differential gene expression (DGE).

Criteria	Annotation Differential	Gene Expression
Biological replicates	Not necessary but can be useful	Essential
Coverage across the transcript	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not as important; however the only reads that can be used are those that are uniquely mappable.
Depth of sequencing	High enough to maximize coverage of rare transcripts and transcriptional isoforms	High enough to infer accurrate statistics
Role of sequencing depth	Obtain reads that overlap along the length of the transcript	Get enough counts of each transcript such that statistical inferences can be made
DSN	Useful for removing abundant transcripts so that more reads come from rarer transcripts	Not recommended since it can skew counts
Stranded library prep	Important for de Novo transcript assembly and identifying true anti-sense trancripts	Not generally required especially if there is a reference genome
Long reads (>80 bp)	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not generally required especially if there is a reference genome
Paired-end reads	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not important

Annotation

- The goal of annotation is to identify genes and genic architecture by analyzing short-sequence reads derived from expressed RNA. Thus, the most important parameter is that the sequence reads evenly cover each transcript, including both ends.

annotation

- Two of the biggest challenges related to using short-read sequencing for annotation are that the sequence reads are much shorter than the biological transcripts and that in many cases a reference genome is unavailable.

Differential gene expression

- The primary goal of DGE is to quantitatively measure differences in the levels of transcripts between two or more treatments or groups. At one level this is as simple as comparing the counts for reads that come from each transcript. However, simple read counts alone are neither sufficient for identifying differentially expressed genes nor for quantifying the degree of differences.

Differential gene expression

- Several types of variances contribute to RNA-seq data including:

- **Sampling variance**: Even though a sequencing run is capable of producing millions of reads, these represent only a small fraction of the nucleic acid that is actually present in the library. There is therefore a built in sampling variance.
- **Technical variance**: Library preparation and sequencing procedures involve a series of complex chemical reactions which all contribute to between sample variance.
- **Biological variance**: Ultimately we are interested in measuring differences between different biological systems. Biological systems are inherently complex and very sensitive to perturbations. Thus, even in the absence of sampling and technical variance biological variance will always exist.

Replication and DGE

- Decisions about the number and types of replicates (individual units of statistical inference) are driven by both extrinsic and intrinsic factors.
- Extrinsic factors include cost, availability of samples, and feasibility of experiments.
- Intrinsic factors are often more difficult to grasp without prior information about the system and definitely more ambiguous from a decision standpoint.

- Unless you are genuinely interested in comparing technical aspects of RNA-seq, or you expect technical variation to be especially great for a large majority of the target transcripts, I **recommend greater resource allocation to biological replication.**
- This isn't to say technical replicates should never be sequenced, but simply that limited financial resources are probably better spent on more biological, relative to more technical, replicates in many cases.

How many replicates?

- As with any experiment that is intended to test a null hypothesis of no difference between or among groups of individuals, differential expression studies using RNA-seq **data need to be replicated in order to estimate within- and among-group variation.**

- Failure to reject the null hypothesis of no difference when there actually is a difference (a "false negative") is known as **type II error**. The number of replicates per group in an experiment directly affects type II error.

How many replicates?

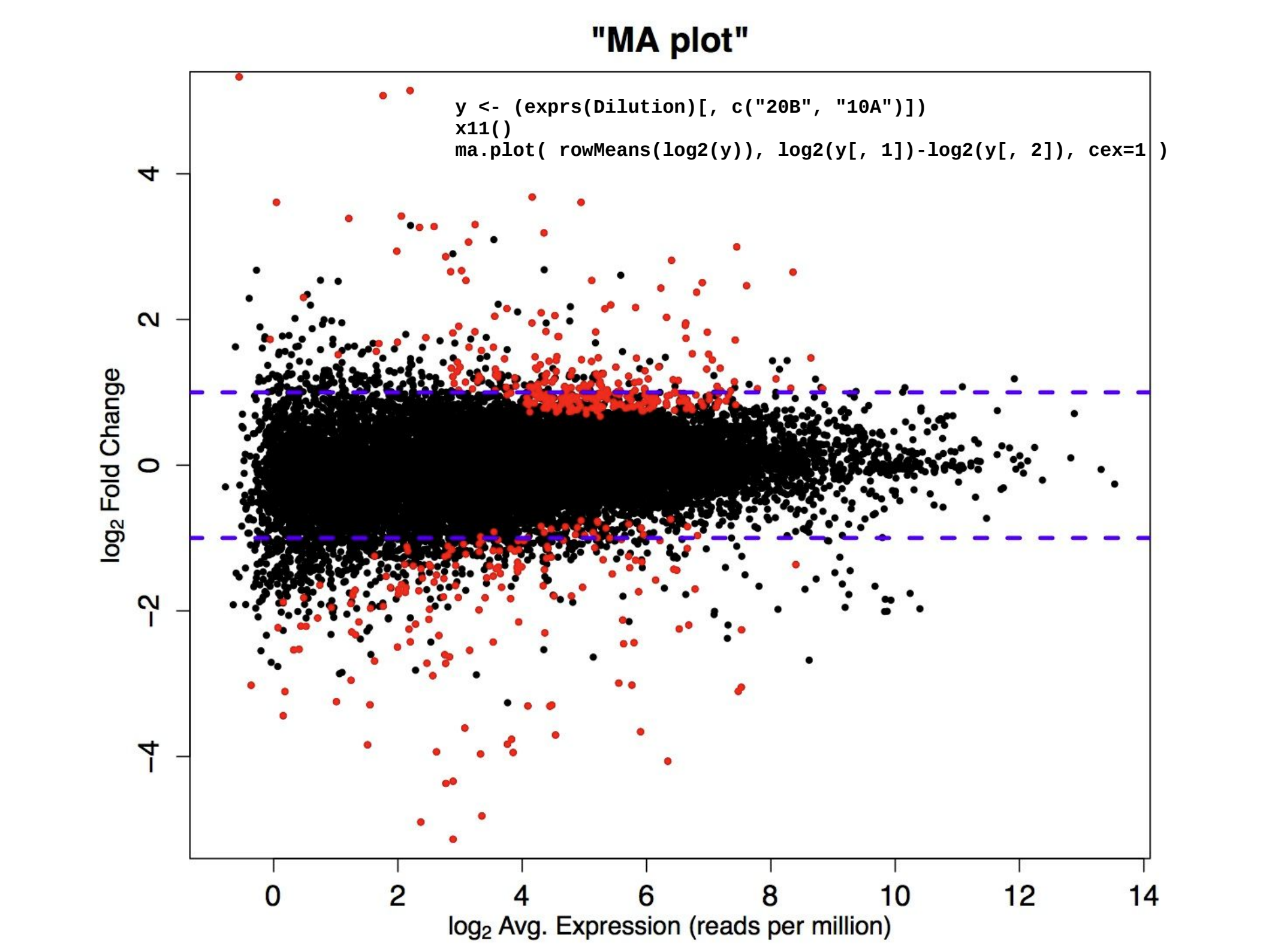
- Power depends on the acceptable maximum probability of **type I error** (the event in which the null hypothesis is rejected in favor of the alternative when the null hypothesis is actually true).
- Experimenters conventionally tolerate the risk of type I error (symbolized by α) if it is less than 0.05 for a given test. When performing many hypothesis tests (as is the case for differential expression tests across thousands of transcripts), type I error must be considered in light of the many tests.

Noble, WS. (2009). How does multiple testing correction work?. Nat Biotech 27:1135-1137

- While **it is impossible for anyone to say with complete accuracy how many replicates are necessary for any given experimental objective,** there are a few key guidelines relevant to the issues described above that will at least help you avoid conducting a completely ineffective study.

Depth of sequencing

- As mentioned previously, variation due to the sampling process makes an especially large contribution to the total variance among individuals for transcripts represented by few reads.
- This means that identifying a treatment effect on genes with shallow coverage is not likely amidst the high sampling noise


"MA plot"
y <- (exprs(Dilution)[, c("20B", "10A")])
x11()
ma.plot(rowMeans(log2(y)), log2(y[, 1])-log2(y[, 2]), cex=1)
log2 Fold Change
4
2
0
−2
−4
0
2
4
6
8
10
12
14
log2 Avg. Expression (reads per million)

Experimental complexity

- A comparison between two groups (e.g. one "experimental" and one "control") is the simplest type of differential expression study design, and probably a fairly common one.

For more complex designs, however, it is important to remember that sufficient replication has to occur at every level of comparison.

- A hypothetical comparison of a simple "one-factor, two-treatment" experimental design with a factorial "three-factor, two-treatment" design. The result in terms of total sample size for 6 replicates per level is $2^1*6 = 12$ versus $2^3*6 = 48$.

Drug 1 with 2 treatments (+ or -)

+ N = 6	- N = 6

Total # ind. = 12

Drug 1 with 2 treatments (+ or -) by Drug 2 with 2 treatments (+ or -) by Drug 3 with 2 treatments (+ or -)

- / - / - N = 6	- / - / + N = 6	- / + / - N = 6	- / + / + N = 6
+ / - / - N = 6	+ / - / + N = 6	+ / + / - N = 6	+ / + / + N = 6

Total # ind. = 48

Sokal, R.R., and F.J. Rohlf. (1995). Biometry: The principles and practice of statistics in biological research. 3rd edition. W.H. Freeman, New York.

Treatment effect size

- The effect size will to some extent be affected by sampling variance, and therefore by depth of coverage.
- There are also **clear biological reasons** to expect differences in effect size across transcripts or specific types of samples. Some treatments will simply affect transcriptional targets more directly than others, so large differences between treatments for these targets are expected.

Tips

- One strategy, provided the experimental setup is not too costly, might be to run the desired experiment with a substantial level of replication (e.g. 10-20 individuals per treatment or combination of treatments), but only actually sequence 3 or 4 samples at moderate to high coverage (e.g. 10 million reads per individual) for each treatment level.

- This way you would have enough preliminary information to conduct a reliable power analysis, but still have plenty of remaining samples "in the bank" for additional sequencing.

Scotty - Power Analysis for RNA Seq Experiments

Marth Lab

Help

Scotty is a tool to assist in the designing of RNA Seq experiments that have adequate power to detect differential expression at the level required to achieve experimental aims.

At the start of every experiment, someone must ask the question, "How many reads do we need to sequence?" The answer to this question depends on how many of the truely differentially expressed genes need to be detected. A greater number of genes will be found with an increase in the number of replicates and an increase in how deeply each existing replicate is sequenced. These parameters are limited by the budget for performing the experiment.

The power that is available using a given number of reads will differ between experiments. Ideally, pilot runs of your experiment (small runs of at least two replicates from one of your conditions) should be used to assess the amount of biological variance that is in the system you are studying, and the amount of sequencing depth that is required to adequately measure the genes. Alternatively, Scotty can be run on data from publicly-available datasets that are very close to your expected experiment (species, library preparation protocol, sequencing technology, and read length).

The Matlab code that runs background calculations is available on github. Please contact us if your require assistance

Inputs

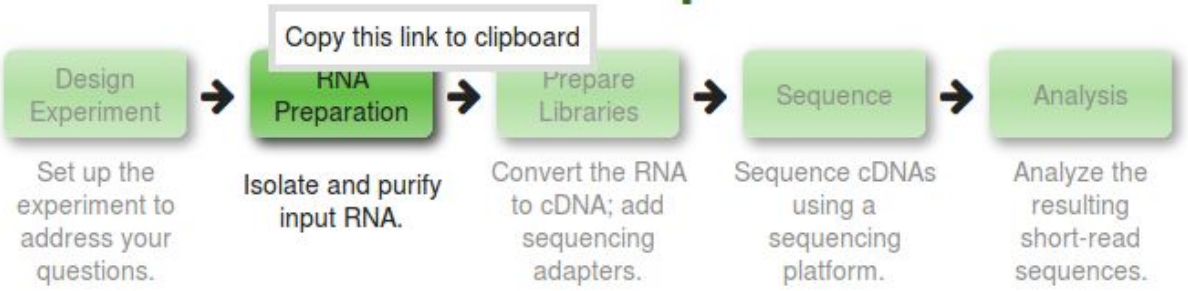
Pilot Data: Upload your own pilot data or used a stored dataset as a model for your experiment. (?)

CAUTION

Power analysis results will not be predictive of the actual results unless the power analysis is performed on data that closely matches the experiment. Please read about generating pilot data and selecting preloaded datasets before continuing.

http://scotty.genetics.utah.edu/scotty.php

2. RNA Preparation



Since the goal of RNA-seq is to characterize the transcriptome the first step naturally involves isolating and purifying cellular RNAs.

Isolation and purification of RNA typically involves disrupting cells in the presence of detergents and chaotropic agents. Depending upon the starting material mechanical disruption is also recommended.

After homogenization, RNA can be recovered and purified from the total cell lysate using either liquid-liquid partitioning or solid-phase extraction.

Working with RNA

- The success of RNA-seq experiments is highly dependent upon recovering pure and intact RNA.
- Because RNA is more labile than DNA and **RNases are very stable enzymes** extra care should be taken when purifying and working with RNA.

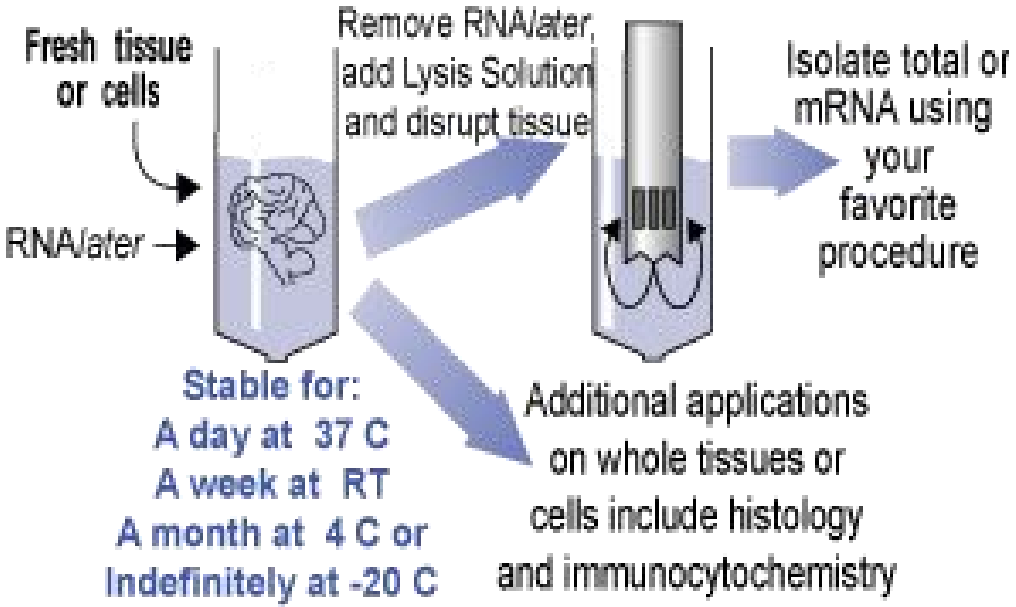
Precautions when working with RNA

1. Maintain separate reagents and consumables for RNA extraction. Avoid co-storage with reagent kits that include RNases.
2. Process the samples quickly and keep the RNA on ice when possible.
3. Wear gloves and work in a clean workspace.
4. Use RNase free water. Water from commercial purifiers is generally RNase-free (as long as the purifier and dispensers are kept thoroughly clean). To ensure that water is RNase-free it can be treated with diethyl pyrocarbonate (DEPC) as follows. Add DEPC to 0.05% and incubate overnight at room temperature. The DEPC is then destroyed by autoclaving for 30 minutes.
5. Glassware can be heated at 250 °C for several hours to remove RNase contamination.
6. Plastic-ware can be soaked in 0.1 N NaOH/1mM EDTA then rinsed thoroughly with RNase free water.

RNA stabilization

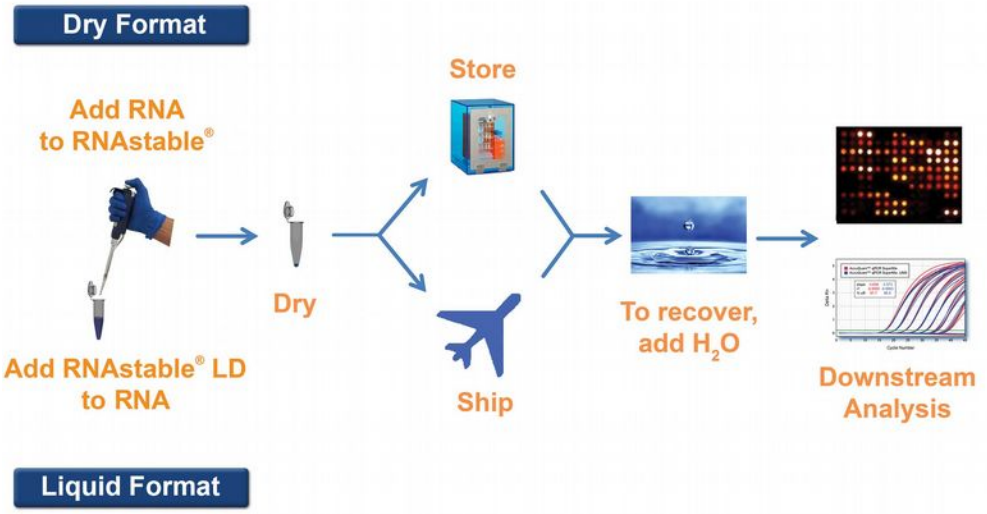
- RNA is much more labile than DNA, and the moment a tissue sample is dissected, for example, the cells begin to die and the RNA begins to be degraded. Since there is often a time delay between harvesting and isolation of the RNA samples should be frozen in liquid nitrogen immediately upon harvest.
- In addition several reagents have been developed to preserve the RNA in fresh tissue. Two of the more commonly used reagents are RNAstable® (from Biomatrica) and RNAlater® (from Qiagen).

RNAlater



- In a beaker, combine 40 ml 0.5 M EDTA, 25 ml 1M Sodium Citrate, 700 gm Ammonium Sulfate and 935 ml of sterile distilled water, stir on a hot plate stirrer on low heat until the Ammonium Sulfate is completely dissolved.
- Allow to cool, adjust the pH of the solution to pH 5.2 using 1M H_2SO_4. Transfer to a screw top bottle and store either at room temperatureor refrigerated.

RNAstable



- RNAstable is based on the natural principles of anhydrobiosis ("life without water"), a biological mechanism employed by organisms such as tardigrades and brine shrimp that enables their survival while dry for up to 120 years. Anhydrobiotic organisms protect their DNA, RNA, proteins, membranes and cellular systems, and can be revived by rehydration.

RNAstable works by forming a glass-like shell, securely "shrink-wrapping" RNA samples and protecting against degradation.

RNA solubilization

- Two of the primary challenges of the isolation procedure are (1) to separate the RNA from other cellular materials and (2) to preserve the integrity of the RNA that is extracted.
- Virtually all RNA extraction procedures rely upon disrupting tissues in an aqueous solution containing organic buffers, guanidinium salts, and ionic detergents.

RNA solubilization

- The buffers that are most commonly used include TRIS and sodium-acetate, which are added to adjust and maintain the pH.
- Chaotropic agents (compounds that disrupt both hydrophobic and hydrogen-bond interactions) are added to denature proteins.
- The most commonly used chaotropic agents are **guanidinium salts**. Guanidinium is a strong protein denaturant capable of denaturing recalcitrant proteins such as RNases.

RNA solubilization

- The benefits of using guanidinium for purification of RNA were first reported in 1951 (Volkin and Carter 1951) and since that time virtually all RNA extraction procedures have incorporated the use of h**igh concentrations (4-6M) of guanidinium thiocyanate or guanidinium hydrochloride**.

- Ionic detergents are added to help solubilize cell membranes and lipids. The most commonly used detergents include sodium dodecyl sulfate (SDS), sodium lauroyl sarcosinate (sarkosyl), sodium deoxycholate, and cetyltrimethylammonium bromide (CTAB).

- Mechanical Homogenization
-

Tissue specific considerations

Muscle and skin:
Isolation of RNA from muscle and skin tissues can be difficult due to the presence of connective tissue, collagen, and contractile proteins. For these tissues proteinase K can be added to degrade the problematic proteins. Proteinase K is an especially robust enzyme that retains activity under conditions that denature most other proteins. For these samples it is generally advised to first use mechanical homogenization in high concentrations of guanidinium. Then the homogenate should be diluted to decrease the guanidinium concentrations to ~2M before adding the proteinase K. However, researchers should be aware this dilution step can affect the downstream purification and in all cases they should follow protocols that have been developed specifically for these tissues.

Tissues high in fat:
Homogenates from tissues that are high in fat (*e.g.*, brain, adipose tissue, and plant seeds) should be extracted with chloroform to remove lipids.

- organic extraction

Considerations when using phenol/chloroform partitioning

1. One of the most common mistakes is to start with too much tissue, since the pH, and the concentrations of guanidinium, phenol, and chloroform are all critical factors.
2. Care must be taken to ensure that none of the interphase or organic phase is withdrawn during removal of the aqueous layer. It is better to sacrifice some of the aqueous layer than to try to recover all of it. If it is very important to recover all of the RNA then it is better to recover a safe amount of the initial aqueous layer then add more extraction buffer and perform a second extraction. The two aqueous layers can then be combined and ethanol precipitated.
3. After the phenol/chloroform extraction it is often beneficial to perform a final extraction using just chloroform to remove traces of phenol, which can inhibit downstream reactions.

- solid-phase extraction
  - In the 1970's it was demonstrated that nucleic acids in the presence of high concentrations of chaotropic agents would bind to silica. Since nucleic acids and silica are both negatively charged the interaction is not simply due to charge-charge interactions. Instead, binding is thought to be due to the formation of a cationic salt bridge, which is promoted by chaotropic compounds.

Solid-phase extraction vs. organic extraction

Both techniques are commonly used. However each has certain advantages and disadvantages. If appropriate care is taken both techniques are capable of producing highly purified RNA.

- Organic extraction is cheaper and is easier to scale up for larger amounts of starting material.
- Silica columns are more expensive but are easier to use and are more amenable for processing multiple samples in parallel.
- Solid-phase extraction has the advantage that the RNA can be eluted using very small amounts of elution buffer, which obviates the need to perform an ethanol precipitation step.

In practice many protocols actually combine both procedures. In this case, after adding appropriate amounts of alcohol to the aqueous phase from the phenol/chloroform extraction, it is further purified using a silica column.

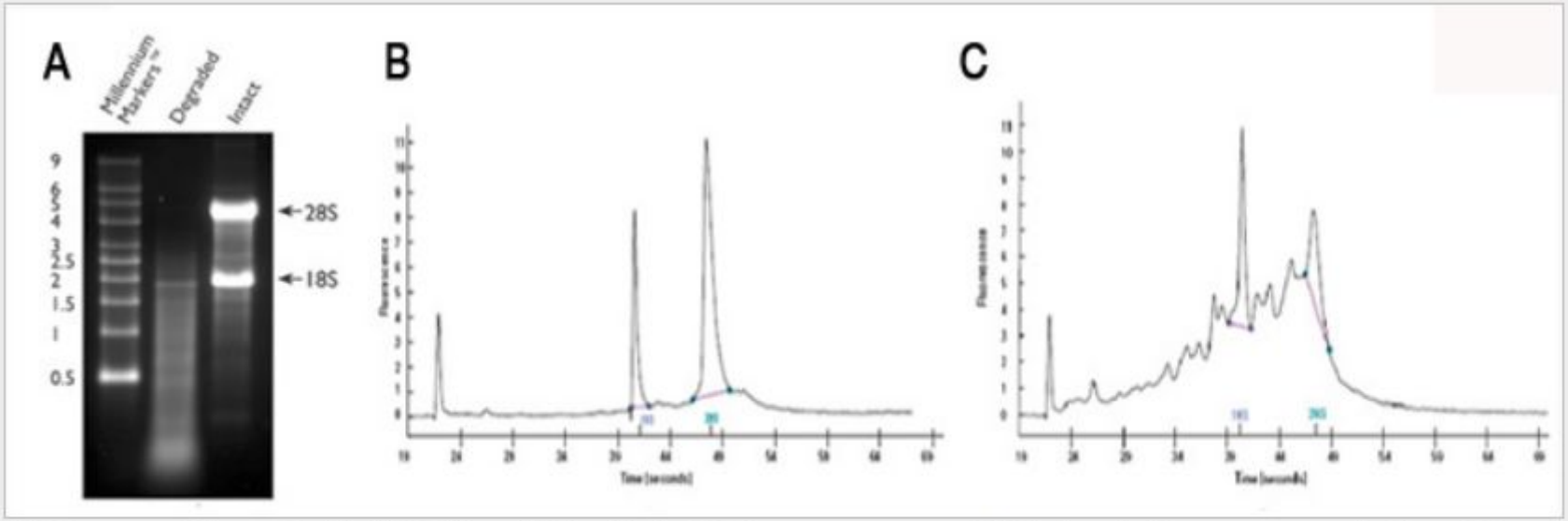
Small RNAs

It should be noted that protocols for isolating mRNA are not generally optimized for isolating small RNAs such as siRNA, piwiRNA, and miRNAs. If the researcher is interested in purifying these RNAs then they must be sure to use methods that have been specifically designed for small RNAs.

Removing genomic DNA contaminants

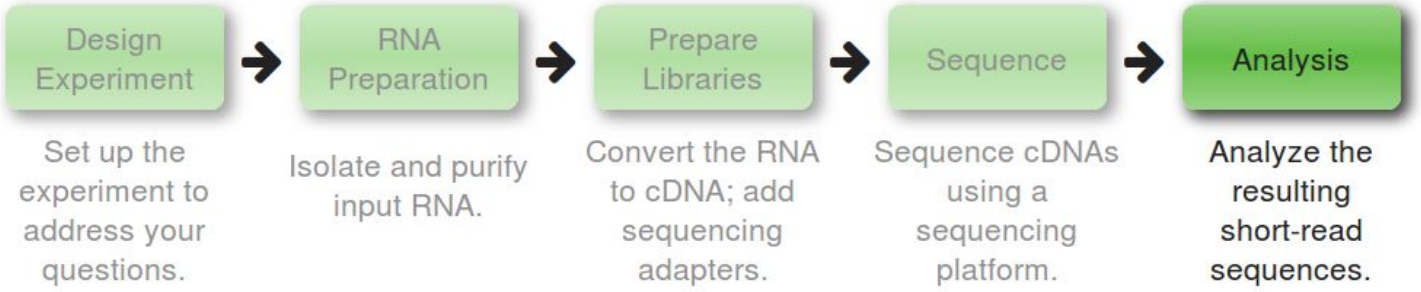
Generally, if one carefully follows established protocols very little genomic DNA is carried over into the final total RNA sample. However, it is sometimes necessary to digest DNA using DNase I.

Assessing RNA quality and quantity



(A) Two µg of degraded total RNA and intact total RNA were run on a 1.5% denaturing agarose gel. The 18S and 28S ribosomal RNA bands are clearly visible in the intact RNA sample. The degraded RNA appears as a lower molecular weight smear. Image from www.ambion.com Panels (B and C) show Bioanalyzer traces of intact (B) and degraded (C) eukaryotic RNA (B) There are two well-defined peaks corresponding to the 18S and 28S ribosomal subunits and the ratio between the 28S and 18S peaks is approximately 2:1. (C) is an example of partially degraded RNA. The 2:1 ratio between the ribosomal peaks is absent and there is a high presence of degraded products.

5. Analysis



Stereotypical RNA-seq Analysis Pipeline

1. Demultiplex, filter, and trim sequencing reads.
2. Normalize sequencing reads (if performing *de novo* assembly)
3. *de novo* assembly of transcripts (if a reference genome is not available)
4. Map (align) sequencing reads to reference genome or transcriptome
5. Annotate transcripts assembled or to which reads have been mapped
6. Count mapped reads to estimate transcript abundance
7. Perform statistical analysis to identify differential expression (or differential splicing) among samples or treatments
8. Perform multivariate statistical analysis/visualization to assess transcriptome-wide differences among samples

Initial processing of raw reads

- Before either de novo assembly or alignment of RNA-seq reads to a reference, they must be processed for a number of reasons described in this section.
- Outline:
  - Demultiplex by index or barcode
  - Remove adapter sequences
  - Trim reads by quality
  - Discard reads by quality/ambiguity
  - Filter reads by k-mer coverage
  - Normalize k-mer coverage

Filtering/Trimming reads by quality

- The sequencing reads, along with the corresponding base call qualities are delivered to the user typically as a FASTQ file (which has the extension ".fastq" or ".fq"). These FASTQ files contain a four-line record for each read, including its nucleotide sequence, a "+" sign separator (optionally with the read identifier repeated), and a corresponding ASCII string of quality characters.

@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!'"*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC65

- Each ASCII character corresponds to an integer **i** ranging from -5 to 41 (depending on the version of Illumina software used for base-calling), and may be translated to p, the probability that a given base is incorrectly called, using either the "Phred" scale where $\mathbf{i = -10*log_{10}(p)}$, or a modified scale where $i = -10*\log_{10}[p/(1-p)]$, again, depending on the version of base-calling software used by the sequencer.

De novo Transcriptome Assembly

- If you can't align RNA-seq data to a well-assembled reference genome from a relatively recently diverged organism, building a reference transcriptome out of RNA-seq reads is an alternative strategy.
- The primary goal in assembling a transcriptome de novo is to reconstruct a set of contiguous sequences (contigs) presumed to reflect accurately a large portion of the RNAs actually transcribed in the cells of your organism.

Aligning RNA-seq reads to a reference

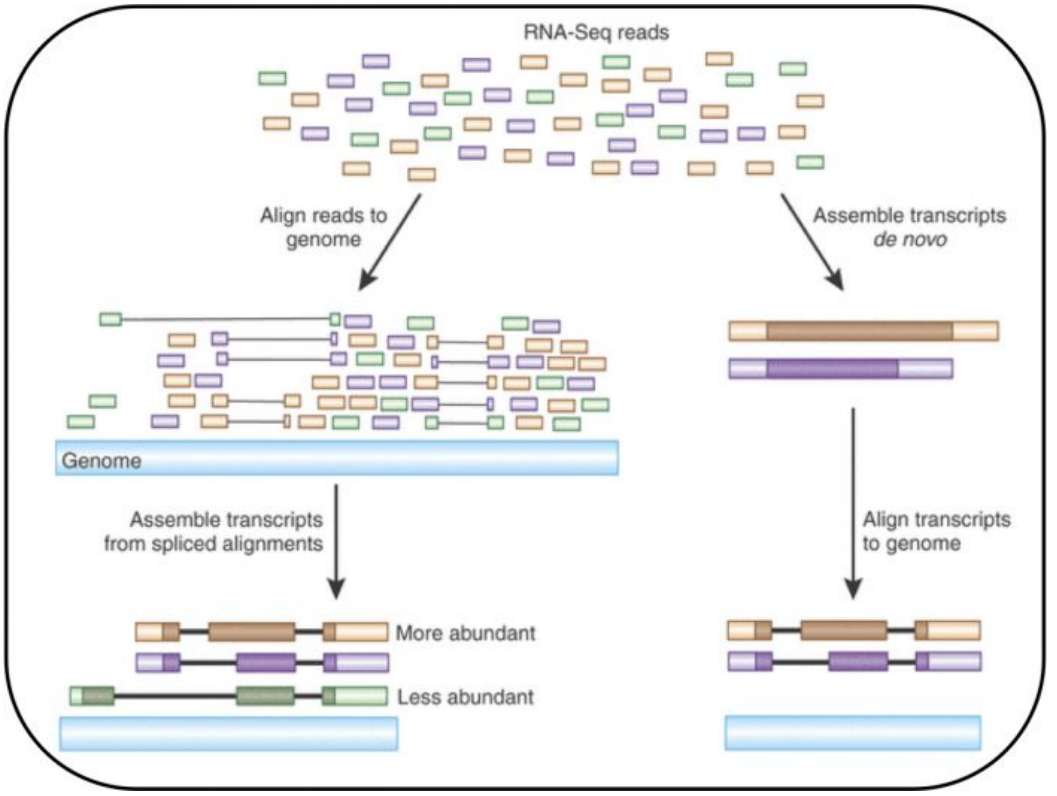
- Short reads generated by RNA-seq experiments must ultimately be aligned, or "mapped" to a reference genome or transcriptome assembly. Read alignment to a reference provides biological information in two basic ways.
- First, it generates a dictionary of the genomic features represented in each RNA-seq sample. That is, aligned reads become annotated, and the highly fragmented data is thus connected to the gene families, individual transcripts, small RNAs, or individual exons encompassed by the original tissue sample.
-

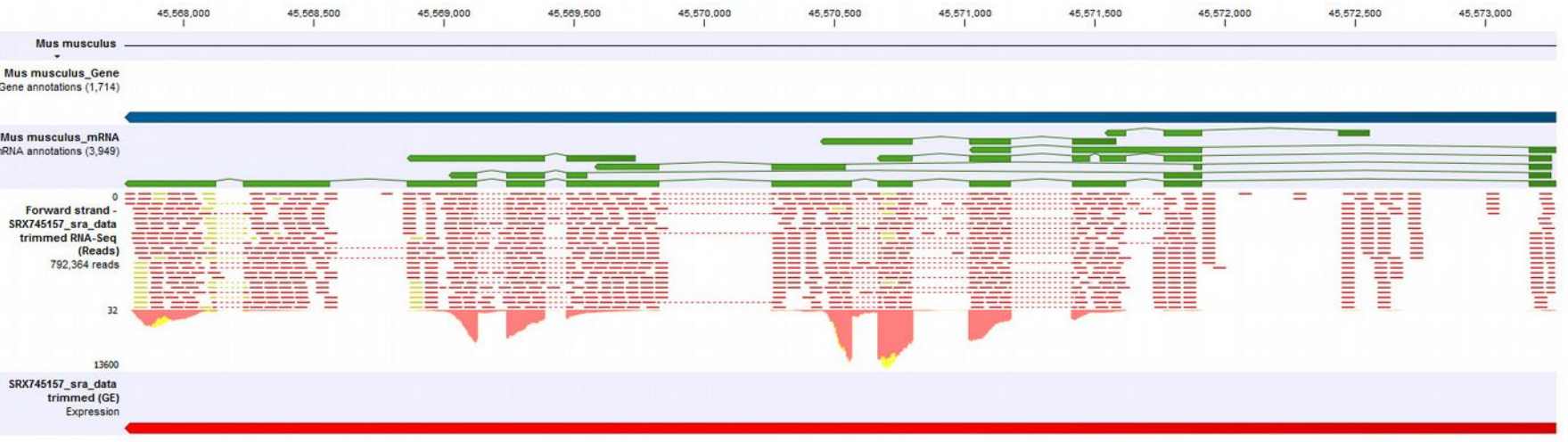
Aligning RNA-seq reads to a reference

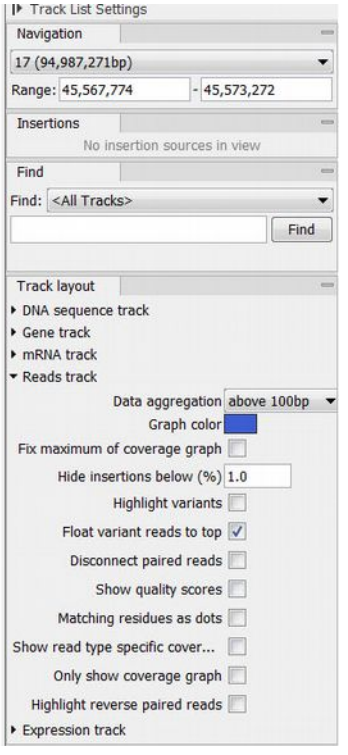
- Second, **the number of reads aligned to each feature approximates abundances of those features in the original sample**. Such measures of digital gene expression are then subject to comparison among samples or treatments in a statistical framework.

Aligning RNA-seq reads to a reference genome

- If you are mapping your reads to a reference genome, you need an alignment tool that can allow for a read to be "split" between distant regions of the reference in the event that the read spans two exons.
-
- Since the sequencing library was constructed from transcribed RNA, intronic sequence was not present, and the sequenced molecule natively spanned exon boundaries. When aligning back to a reference genome we have to account for reads that may be split by potentially thousands of bases of intronic sequence.


RNA-Seq reads
Align reads to genome
Assemble transcripts de novo
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant


45,568,000
45,568,500
45,569,000
45,569,500
45,570,000
45,570,500
45,571,000
45,571,500
45,572,000
45,572,500
45,573,000
Mus musculus
Mus musculus_Gene
Gene annotations (1,714)
Mus musculus_mRNA
RNA annotations (3,949)
0
Forward strand - SRX745157_sra_data trimmed RNA-Seq (Reads)
792,364 reads
32
13600
SRX745157_sra_data trimmed (GE)
Expression


Track List Settings
Navigation
17 (94,987,271bp)
Range: 45,567,774 - 45,573,272
Insertions
No insertion sources in view
Find
Find: <All Tracks>
Find
Track layout
DNA sequence track
Gene track
mRNA track
Reads track
Data aggregation above 100bp
Graph color
Fix maximum of coverage graph
Hide insertions below (%) 1.0
Highlight variants
Float variant reads to top
Disconnect paired reads
Show quality scores
Matching residues as dots
Show read type specific cover...
Only show coverage graph
Highlight reverse paired reads
Expression track

microRNA aligners

- Some aligners are designed to map especially short sequences, such as those obtained from mircoRNA-seq studies. These aligners include parameter value ranges necessary for the optimal alignment of reads that are only 16-30 nt.
-