Verma et al. BMC Genomics (2025) 26:918
https://doi.org/10.1186/s12864-025-12094-8

BMC Genomics

REVIEW

Open Access

# Commentary: a review of technical considerations for planning an RNA-Sequencing experiment

Rashi Verma[1,2], Amanda Savaria-Butler[3], Francisco J. Enguita[4] and Robert Meller[1,2*]

**Abstract**

There are a bewildering number of variables to consider when planning an RNA sequencing study (RNA-Seq). Here we discuss some of the considerations for an investigator initiating such studies, with a focus on library preparation, depletion techniques, considerations for data analysis, and finally, we consider the cost of such a study. The goal of this summary is to make the researcher aware of such considerations to be able to plan an analysis strategy prior to initiating an RNA-Seq study.

**Keywords** RNA-Sequencing (RNA-Seq), Transcriptomics, Experimental Design, Data Analysis, Technical Considerations

## Introduction

Since RNA-Sequencing (RNA-Seq) became an accepted method to assess RNA expression in cells [1], many studies have utilized this approach to assess RNA expression and transcript usage, as well as miRNA expression and the characterization of non-coding RNAs. It is critical to acknowledge that RNA-seq does not count absolute numbers of RNA copies in a sample, rather it yields relative expression within a sample of RNA (see below). Although gene expression measurements are highly consistent between RNA-Seq and microarray approaches, the appeal of RNA-Seq over microarray approaches is the ability to sequence, quantify novel RNA species, and assess alternative splicing, which may have relevance for normal and disease states [2]. Furthermore, unlike microarrays, RNA-Seq is not limited by unequal labeling efficiency of fluorescent dyes, the risk of non-specific binding due to cross-hybridization of sequences with high identity, or dynamic range limitation of the scanner, which makes detection of both low and high expressing genes simultaneously impossible. However, for the uninitiated RNA-Seq studies ask as many questions as it may answer. Therefore, this commentary raises considerations and approaches for performing RNA-Seq studies based on our experiences with blood [3–6].

*Correspondence:
Robert Meller
rmeller@msm.edu
[1]Morehouse School of Medicine, Institute of Translational Genomic Medicine, Atlanta, GA 30310, USA
[2]Morehouse School of Medicine, Neuroscience Institute, Atlanta, GA 30310, USA
[3]Space Biosciences Division, NASA Ames Research Center, Amentum, Moffett Field, CA, USA
[4]Faculdade de Medicina Universidade de Lisboa, Av. Prof. Egas Moniz, Lisbon 1649-028, Portugal

## Basic considerations of experimental design

### Study design

First, ascertain what biological question you wish to answer with your study. A clear experimental question, study design, and hypothesis will help in designing the subsequent statistical analysis of the data you will generate. One should be realistic about the implications and interpretation of a study, for example a study of

RNA expression in blood is just that. While some gene expression changes in blood may reflect patterns of gene expression changes in other tissues [7], one should guard against over interpretation. Similarly, one must accept the limitation of the assessed tissue/biosample in a study looking for surrogate biomarkers [4].

### Biotype of RNA

The second consideration is which RNA do you want to measure? Messenger RNAs (mRNA) encode proteins. In contrast, non-coding RNAs, such as long non-coding RNAs (lncs) whilst not encoding for proteins, have biological properties in themselves either as structures interacting with other proteins or RNAs [8]. One class of non-coding RNAs are micro RNAs (miRNAs), which are best known as regulators of RNA stability, via the regulation of RISC complex mediated RNA degradation [9–12]. Some novel RNAs may appear in sequencing samples [13, 14], in intergenic and intronic gene regions, but have yet to be annotated in the reference genome. Some of these novel RNAs, may be associated with clinically relevant SNPs annotated in the clinVar database (https://www.ncbi.nlm.nih.gov/clinvar/) (for an example see ref [4]). Finally, other RNA families such circular RNAs [15], nuclear RNAs, and other small RNAs each have their own unique biology, and may need specific design for the sequencing experiment.

It should be noted that not all RNAs are assessed by all techniques. For example, using probe-based detection methods such as Nanostring, Amplicon sequencing, or Microarray will only detect the RNA of interest if that assay has probes specific to the gene of interest. Small RNAs, including miRNAs, are usually removed from standard mRNA-Seq workflows when size selecting, and thus require specialized assays and library preparation for detection. Larger non-coding RNAs may not be polyadenylated, and as such require whole transcriptome library building workflows (see below). However, most standard sequencing platforms (Illumina, O.N.T., etc.#[1]) have standardized library building workflows for the sequencing of mRNA, as well as small RNAs, and non-polyadenylated RNAs.

## Library building considerations

The first step of preparing RNA for sequencing is to build a library from the sample of extracted RNA. This typically involves reverse transcription of the RNA into complementary DNA (cDNA), the addition of sequencing adapters, followed by PCR-based amplification of the library, depending on the requirements of the sequencing platform.

[1] not an endorsement of any platform

### RNA quality

RNA quality is not something that can be easily fixed once compromised, which is why attention to detail at every step of the workflow is crucial. From collection to extraction and quality check (QC), every stage must be carefully considered to ensure that the RNA used for sequencing is of the highest quality possible. The integrity of RNA directly affects the accuracy and depth of transcriptomic analysis, as degraded RNA can lead to biases, particularly in the detection of longer transcripts or low-abundance genes. A commonly used measure for evaluating the quality of RNA is the RNA Integrity Number (RIN), a quantitative relationship between the amount of ribosomal RNA species; values greater than 7 generally indicate enough integrity for high-quality sequencing [16, 17], but this general rule of thumb may vary depending on the source of the biological sample that could prevent the obtaining of high quality RNA. However, blood samples often present challenges in maintaining high RNA integrity, requiring extra attention during sample collection and handling. Blood should ideally be collected using RNA-stabilizing reagents, such as PAXgene®, or processed immediately to isolate cells or plasma, followed by storage at −80°C to preserve RNA from degradation. Upon extraction, it is crucial to perform rigorous QC checks. In addition to measuring RIN, the 260/280 and 260/230 ratios should be assessed to ensure minimal contamination of protein or DNA during the extraction phase. Electropherograms generated by systems Like Bioanalyzer or TapeStation can visually confirm RNA integrity, with a healthy sample showing distinct 28S and 18S rRNA peaks in a 2:1 ratio. In cases where RNA integrity is compromised, RNA-Seq approaches that capture mRNA by targeting the poly(A) region using Oligo dT beads, and thus require mRNAs to be intact, are not suitable [16, 17]. On comparison, alternative methods that utilize random priming and include steps like ribosomal RNA (rRNA) depletion can enhance performance significantly with degraded samples because they do not depend on an intact polyA tail. Some library building protocols appear better when working with degraded RNA samples [18], and rRNA depletion can be beneficial as well [19, 20]. Ultimately, RNA quality must be prioritized throughout the entire workflow to ensure that sequencing results are accurate and truly reflective of the biological transcriptome. High-quality RNA is vital for uncovering meaningful insights in blood transcriptomics.

### To strand or not to strand!!

One of the first major differences between RNA-Seq platforms when the technology was evolving was whether to create a strand aware Library. The two considerations for this are 1) whether you are interested in determining if an RNA was transcribed from the + or – strand of

DNA. This is critical when trying to identify novel RNAs or RNAs that overlap on opposite strands of the genome, or for the determination of expression isoforms generated by alternative splicing. 2) Cost, protocol complexity, and the amount of RNA needed as an input (25ng-1μg). Because the unstranded protocols are simpler they are usually cheaper and allow for a lower RNA input (Fig. 1). Some manufacturers were early adopters of a stranded library building approach (which influenced early platform decisions). Early stranded libraries used a direct ligation of adapters to fragmented RNA to achieve a directional library. The subsequent libraries were amplified using primers complementary to unique sequences in the 5’ and 3‘ adapters. More recent approaches use a strand switch protocol using UTP in the second strand generation, followed by degradation of the second strand once adapters are ligated using Uracil-DNA-glycosylase (Fig. 1).

*-Stranded libraries are preferred for the better preservation of transcript information for transcript orientation and long non-coding RNAs.*

### Ribosomal depletion

Approximately 80% of cellular RNA is ribosomal RNA [21]. Human rRNA is encoded by regions of chromosome 13, 14, 15, 21, and 22. In addition, there are two “misplaced contigs” GL000220.1 and KI 270733.1 that contain rRNA sequences. Since most researchers have questions about non-rRNAs, if these rRNAs are sequenced 80% of the data will be associated with rRNA. The consequence of this is an increase in cost to obtain sufficient reads to map to non-rRNA regions of interest. Therefore, to reduce the cost of generating sufficient non-ribosomal data, efforts have been made to deplete samples of rRNA prior to library building. Earlier approaches used rRNA-targeted DNA probes conjugated to precipitating beads

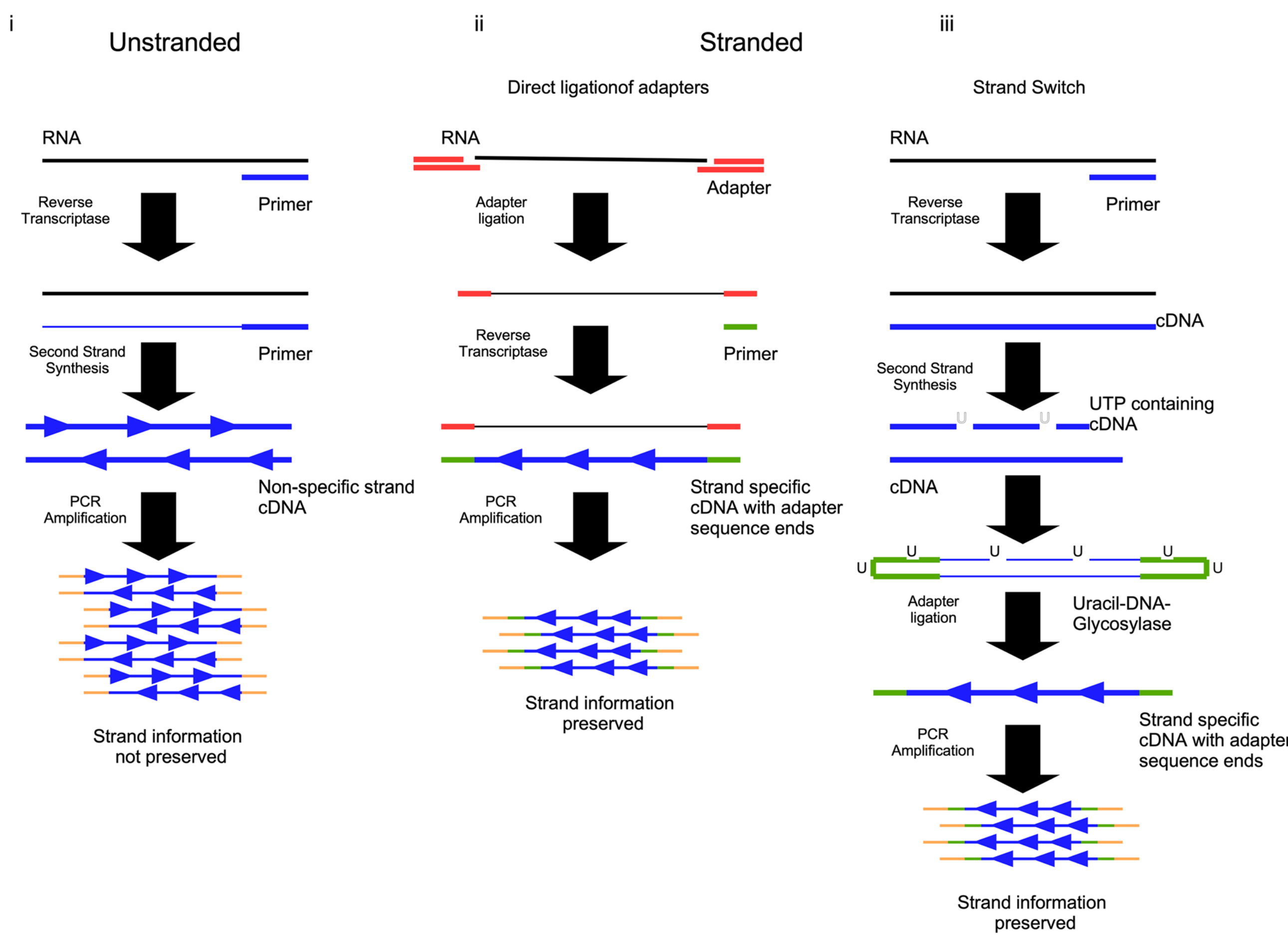


**Fig. 1** RNA-Seq library building methods, using unstranded or stranded approaches. The basic principle of building an RNA seq library is to reverse transcribe fragmented RNA into cDNA, create a second strand, and then amplify using PCR. i) For an unstranded library the PCR reaction is performed following ligation of non-directional adapters containing primer sequences specific for a given platform. ii) Applied Biosystems released a stranded RNA-Seq Library building assay, whereby following fragmentation adapters are Ligated directly to the RNA, and the 3’ adapter was used for priming the RT step. The adapters also include sequences complementary to primers for barcoding and amplification. iii) A second approach utilizes a hexamer RT step of fragmented RNA, followed by a polymerase-mediated second strand synthesis using UTP. The adapters are ligated to the cDNA hybrid, and then the second strand is degraded by a UTP glycosylase, leaving the first strand available for targeted PCR

(magnetic). More recent approaches first hybridize rRNA to complementary DNA probes creating an rRNA-DNA complex then use RNAseH mediated degradation of the RNA-DNA complexes. Once rRNA is depleted, the remaining RNA is sequenced allowing a more cost-effective RNA-Sequencing experiment (Fig. 2).

However, whilst a cost-effective strategy to increase non-rRNA content of a library, there are limitations and considerations for the use of this approach. The first consideration is that depletion is an additional step, and some protocols appear more reliable and reproducible than others. For example, our early assessment of precipitating bead vs. RNAseH based approaches determined that precipitating bead methods gave more effective enrichment of non-ribosomal RNAs but with greater variability (Fig. 2 A and B). In contrast, the relative enrichment of the RNAseH method whilst more modest was more reproducible. The impact of depletion is that relative expression of the gene is enhanced; however, some genes may also be depleted along with the rRNA due to off-target effects. In our assessment we determined differential changes in normalized RNA expression following rRNA removal (Fig. 2). The majority of genes show increased expression (such as when measured by reads per kilobase of transcript per million aligned reads: RPKM values) (Fig. 2 C: STRADB) but some RNAs show decreased levels following rRNA removal (Fig. 2 C: ATG101).

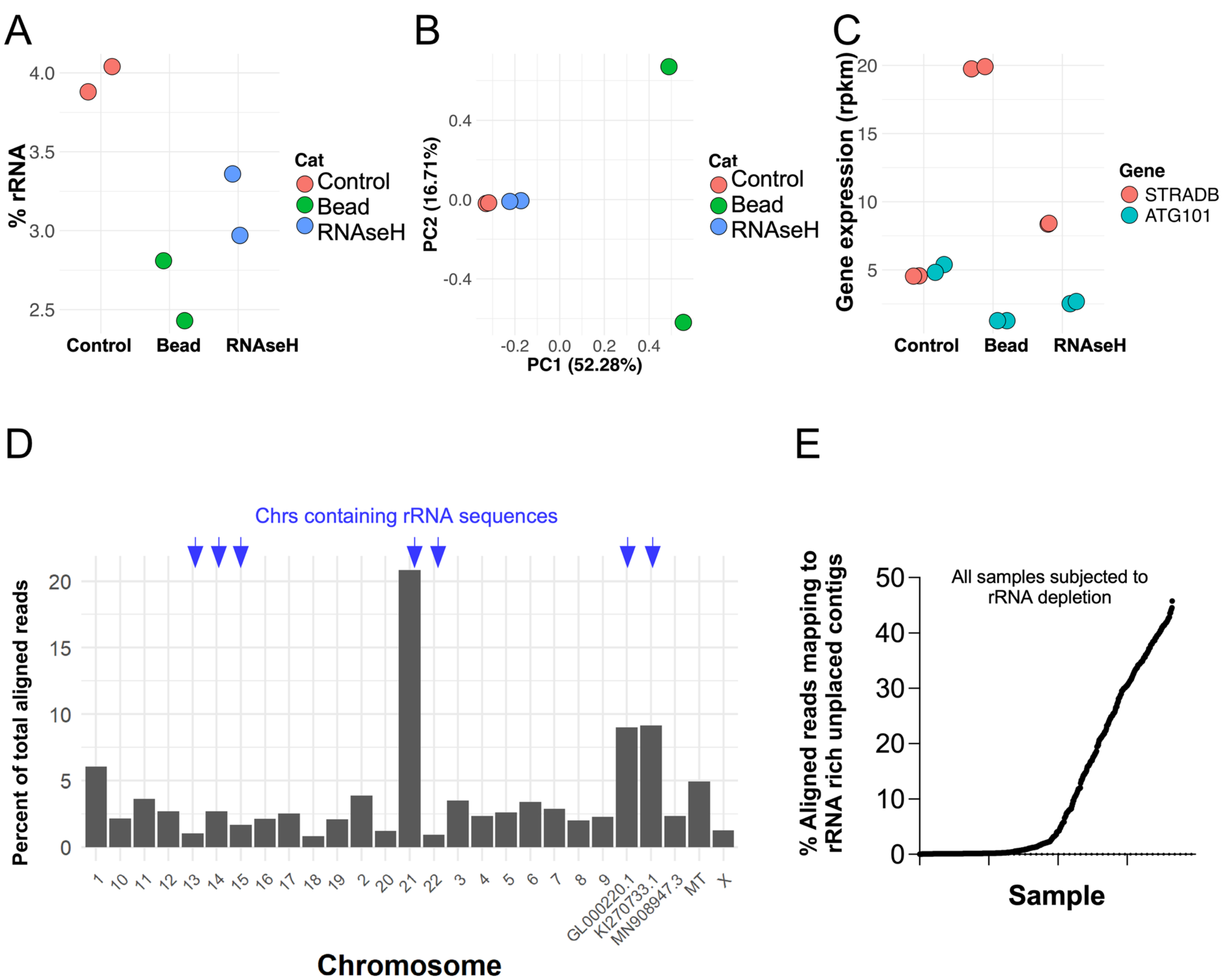


**Fig. 2** Effect of rRNA depletion on gene expression and library variability. **A**-**C** Analysis of RNA-Seq data comparing effect of precipitating bead (Riboclear), RNAseH (NEB) and non-depletion protocols (control) on RNA-Seq library performance. **A**) % of rRNA in samples using Ion Torrent RNA-Seq workflow (NB separation due to jitterplot setting). **B**) Principal component analysis of RPKM expression values for ref-seq identified RNAs. **C**) Individual RPKM values of STRADB and ATG101 gene expression in Control, Precipitating bead rRNA depleted, and RNAseH depletion methods. Note decrease in ATG101 expression following both rRNA depletion methods. **D**-**E**) rRNA analysis of RNA-Seq data presented in [23]. All samples were subjected to rRNA depletion followed by alignment to GRCH38 using STAR. **D**) Percent of total aligned reads mapping to each chromosome of a single RNA-Seq sample with high reads aligned to GL000220.1 from [23]. Chromosomes with rRNA sequences are marked with a blue arrow. **E**) Percent of aligned reads mapping to rRNA located on contigs GL000220.1 and KI270733.1 across 600+ samples (arranged from low to high percent reads mapping to the two contigs)

Therefore, prior to initiating a study, a full assessment of a depletion strategy on any genes of interest is worthwhile.

A second consideration is that the depleted genes cannot be studied. A good example of this challenge is highlighted by globin depletion, which is sometimes performed following rRNA depletion of blood [22]. Since globin genes are a larger component of blood transcriptomes, removal of globin genes as well as rRNA will enhance the number of reads associated with non-globin RNAs. However, we have shown globin genes are regulated following seizures [3], and globin depletion for any study around sickle cell disease for example would be non-sensical, as these genes are highly regulated in this disorder [4].

Finally, even in the hands of experienced labs, rRNA depletion may be highly variable. In a recent COVID-19 study, we noticed that the amount of reads aligning to the two misplaced contigs GL000220.1 and KI 270733.1, which encode rRNAs, was very variable between samples, sometimes accounting for 50% of reads [23] (Fig. 2 D and E). Further investigation of the proportion of reads aligning to these regions suggested that the depletion of rRNA was highly variable between samples (but was not batch related) (Fig. 2 E). We therefore scrubbed the data of rRNA-mapped reads prior to aligning to the reference genome for subsequent gene count quantification [23]. The resultant effect for either biochemical or bioinformatic rRNA depletion is that while individual counts to a specific gene are very similar with and without depletion, the relative counts (CPM or RPKM) are much higher, as the rRNA counts are included in the denominator of the CPM and RPKM calculations.

*-Recommendation: Test each depletion method for reproducibility in a pilot study, and to scrub data of rRNA reads prior to alignment.*

### Poly-A RNA-Seq vs. whole transcriptome

Another library building consideration would be whether to use a poly-A primed library construction or a whole transcriptome approach. Poly-A priming a library restricts the RNA sequencing to only those RNAs with a poly-adenylated tail, yielding a less complex library. However, many RNAs are not subjected to poly-adenylation. The whole transcriptome library building approach uses random hexamer priming of fragmented RNA or direct ligation of adapters for priming to the target RNA (Fig. 1). The challenge downstream for whole transcriptome analysis is that they typically have a large component that is ribosomal RNA, hence depletion strategies are usually considered (see above). Whole transcriptome approaches can detect both immature and mature microRNAs, and other small nuclear RNAs [4]. In addition, novel RNAs including novel non-coding RNA can be identified from intronic and intergenic regions. In our recent study we showed that a number of presumed non-coding clinical variants are present in novel RNAs, which were not present in the reference annotations, and some appeared to have clinical significance [4]. These novel RNAs are easily annotated and quantifiable using programs such as Cufflinks or Stringtie2 [24–26].

*-Recommendation: Future studies should be whole transcriptome and stranded whenever feasible.*

### Long-read vs. short-read platforms

Historically, most RNA-seq studies have been performed using short-read sequencing platforms (50–400 bp). In recent years the ability to perform long-read sequencing (1-50 kb) has emerged [27]. This technology is typified by long read-cDNA protocols for both PacBio and Oxford Nanopore that can sequence the entire span of a transcript, enabling the sequencing of full-length transcripts, allowing for more accurate characterization of transcript isoforms and complex alternative splicing events that are often difficult to resolve with short reads alone. A comprehensive assessment of five different long and short read sequencing techniques revealed that transcript quantification is sensitive to RNA fragmentation and shorter read length, showing that isoforms that appear as the major isoform in short read RNA-seq data may be due to read fragmentation ([28]). Furthermore, as shown by studies of calcium ion channel isoforms in schizophrenia, RNAs show tissue-specific novel isoform expression, which would be difficult to interpret using a short read approach [29]. Such comprehensive transcript resolution is especially valuable not only for well-annotated human genomes but also for research on species lacking comprehensive reference genomes [30, 31]. In non-model organisms, long-read RNA-seq facilitates *de novo* transcriptome assembly and annotation, allowing an investigator to obtain gene discovery and expression profiling in organisms where references are incomplete or unavailable. Established pipelines are available for de-novo transcriptome assembly and annotation [32]. It should be noted that *de novo* transcriptomes can be generated from short read data too, but the long read data typically provides a more complete assessment of the transcriptome. Additionally, certain long-read approaches, such as direct RNA sequencing, bypass the reverse transcription and amplification steps, thereby removing reverse transcriptase (RT) bias, a common source of error in cDNA-based methods (considered below).

Despite the advantages, long reads typically come with higher error rates and greater costs, which can affect quantification accuracy and experimental scalability. Although computational pipelines for transcript detection are available, not all methods are compatible with long-read data [33–35]. In contrast, short-read sequencing remains highly reliable for most use cases, cost-effective, and well suited for higher throughput applications. Analysis pathways from multiple methods have been developed to be compatible with short-read data and are typically benchmarked accordingly. Additionally, short-read platforms are more suited to degraded or fragmented RNA samples. Therefore, the choice between short-read and long-read sequencing technologies in RNA-seq experiments should be carefully aligned with the specific study objectives, balancing resolution needs, budget constraints, and downstream analyses. Together, these complementary approaches reflect the evolving landscape of RNA-seq technologies, offering researchers versatile tools to capture transcript diversity and enhance biological insights across diverse systems and experimental contexts.

### Direct RNA-Seq vs. cDNA-Seq

Almost all methods discussed here deal with sequencing of PCR-amplified cDNA, which is synthesized by the enzyme reverse transcriptase. Therefore, this is more correctly termed cDNA-Seq. cDNA-Seq must therefore contend with the potential for bias of the reverse transcriptase (RT) enzyme when it synthesizes the cDNA first strand, and amplification bias. One method that sequences RNA directly is direct RNA-Seq from Oxford Nanopore Technologies. The advantage of this method is the lack of RT and amplification bias, and the ability to detect modified RNA bases (e.g. N6-methyladenosine (m6A)). The disadvantages of this approach are cost, ONT per base costs are higher than many platforms, the requirement for higher RNA input amounts and high RNA integrity [36] and the lower base calling accuracy of ONT approaches at this time. Third, the availability and constantly changing (updating/improvements of) ONT products and software may be a challenge for some long-term studies. However, the ability of this approach to sequence whole-length transcripts and generate base specific data currently make it unique in the sequencing field ##[2] [37].

*-We do not offer a recommendation for one platform over another; we advise the reader to be aware of the advantages and disadvantages of both approaches and apply the most suitable technology to address their specific experimental question.*

### Batching and sequencing depth

The way a sample is sequenced will have implications for the quality of data obtained. The first hard guideline is the recommendation for barcoding of samples where possible and multiplexing of samples. This should be performed such that each sequencing run/flow cell contains samples of all study variables. For example, one would not want to run all the controls in one batch and all of the treated samples on a different run, as any technical variation in library assembly and sequencing would be added to the biological variation between samples. The better strategy would be to split the samples such that each flow cell/run contains both control and treated. This becomes a challenge when looking at time course data and multivariate data sets, but in general matching should be attempted. Longitudinal studies are going to be challenging due to library kit updates, base calling software updates by manufacturers, and batch effects. There are bioinformatic approaches to help resolve this issue (see below).

The depth of sequencing for RNA is calculated differently from DNA. Specifically for DNA the number of reads and length of a reads aligning to a given region for a given genome size is termed coverage and guides sequencing runs. RNA has varying levels of expression, so the study design usually defines the target number of reads per sample. To detect gene expression changes in human samples, for the polyA-capture method 25–30 million aligned reads and for the ribo-deplete method 35–50 million aligned reads are used (40–50 million (polyA) or 60–70 (ribo-deplete) may be required to detect low-expressing transcripts), and for isoform expression up to 100 million reads can be used for exploratory studies. The amount of RNA expression in a given sample, is typically defined by the complexity of the cell. For RNA-Seq the read size does not usually impact the number of reads estimation, however a read length of at least 50 bp is recommended to reduce the number of reads mapping to multiple locations on the reference genome. Read length becomes a more significant factor when investigating transcripts or defining the transcriptome of a given cell/tissue (see Sect. 4.2).

*-Recommendation: Pay attention to batching when sequencing multiple samples to ensure all of one condition are not sequenced in the same batch. Most cores will help with this planning.*

## Data analysis considerations

### Bioinformatics and alignment considerations

The first step of data analysis involves the assessment of data quality, followed by the removal of poor-quality bases and adapter sequences introduced in the library building step. The default for base quality is a phred score of 20 (error rate of 0.05), although some direct sequencing may have lower phred scores at this time. Tools such as FastQC provide a detailed snapshot of the data quality by highlighting issues like low base quality scores,

[2] 2025 Roche are announcing a nanopore platform

abnormal GC content, sequence duplication, or adapter contamination. These early indicators help determine whether samples need to be discarded or corrected. This tool can be used with additional packages to assess multiple files at once to assess sequencing of a batch of samples (examples include multiQC [38] or ngsReports [39]).

Once data are assessed, one can remove unwanted adapter sequences, low-quality bases at the ends, and very short reads that could interfere with accurate alignment and quantification using all-in-one tools such as Cutadapt [40] (implemented in the wrapper script TrimGalore (https://github.com/FelixKrueger/TrimGalore) or Trimmomatic [41], or stand-alone tools such a sickle (https://github.com/najoshi/sickle) and scythe (https://github.com/vsbuffalo/scythe). By removing low-quality and/or contaminating sequences and enforcing minimum read length and quality thresholds, trimming effectively filters out noise resulting in leaving a cleaner, more reliable dataset that improves downstream analyses and ultimately enhancing the accuracy of biological insights drawn from the data.

*-Recommendation: There are small differences in output from each trimming software, the authors suggest piloting each before using.*

For the second step, data alignment to the reference genome, there is a bewildering array of options available for the alignment. Earlier options included modifying parameters of DNA alignment tools such as BWA and Bowtie. While accurate and commonly deemed the gold standard, these approaches were not splice aware and would sometimes not give good performance around RNA splice junctions. This issue was solved by the development of splice-aware RNA-Seq alignment tools, of which STAR has been one of the most popular [42]. Indeed, the consensus NASA short read pipeline uses STAR, for example [43]. Some pipelines use a first pass with STAR followed by Bowtie2 alignment of unaligned reads (Ion Torrent analysis suite). STAR has been upgraded to include single cell capability (STAR Solo) and the online documentation is noteworthy. However, STAR is computationally expensive, hence faster methods to map and quantify RNA-Seq data have been developed including the pseudoaligners Salmon and Kallisto, and for graph genome use hisat2 [25, 44, 45]. For these later alignment/mapping tools the analysis of the data file is typically an order of magnitude faster than STAR but requires more computational power for setting up the index used for mapping. Some mapping software, such as Salmon and STAR allow the direct generation of gene count matrix relative to the supplied reference annotation guide (.gtf file) [46]. It should be noted that many aligners generate similar data in terms of the number of counts aligned to a given gene for a given reference genome (Fig. 3).

*-Recommendation: New alignment methods are being developed and released. Before using a tool it's useful to benchmark to another aligner to ensure similar performance.*

The result of an alignment procedure is to generate an alignment map file, which is then parsed by software such as cufflinks or stringtie to generate a gene and/or transcript count matrix. Prior to parsing the data, post alignment assessment of the data can be useful to identify potential issues. One common challenge is the issue of duplicate reads for RNA. Read duplicates can be biological or technical in nature. Biological duplicates are the result of multiple copies of a single transcript that results in two or more identical read sequences, one from each copy of the transcript present in the original sample. Technical duplicates, on the other hand, arise due to a technical artifact of the library preparation or sequencing process. Types of technical duplicates include PCR duplicates, which are created during the final step of library preparation (the library amplification step), sister duplicates, which happen when complement strands of the same library form independent clusters on the flow cell, optical duplicates, which occur when a single cluster is falsely identified as two separate clusters (unique to unpatterned flow cells), or clustering duplicates, which occur when a single library populates two adjacent wells on the flow cell during cluster formation (unique to patterned flow cells). Since biological duplicates are a true reflection of the transcript abundance in a sample, these duplicates should be quantitated and included in downstream analyses. However, technical duplicates do not reflect true biology and may lead to inaccurate results if not removed prior to downstream analyses. Post-alignment assessment of the bam file using samtools and picard tools can identify and remove such duplicated reads, but these tools cannot inherently distinguish between biological and technical duplicates. One solution to this is to incorporate a Unique Molecular Identifier (UMI) into each molecule during Library preparation, prior to PCR amplification, which can be used to assess the nature of duplication events. The UMI is typically a random 10–16 bp barcode such that the length of the UMI is sufficient to ensure that the likelihood of a biological duplicate receiving the same UMI sequence is extremely improbable. Since the UMI is added prior to PCR amplification, technical duplicates will have the same UMIs whereas true biological duplicates will have different UMIs. Therefore, when multiple reads map to the same coordinates, the UMIs can be used to distinguish biological from technical duplicates. The use of

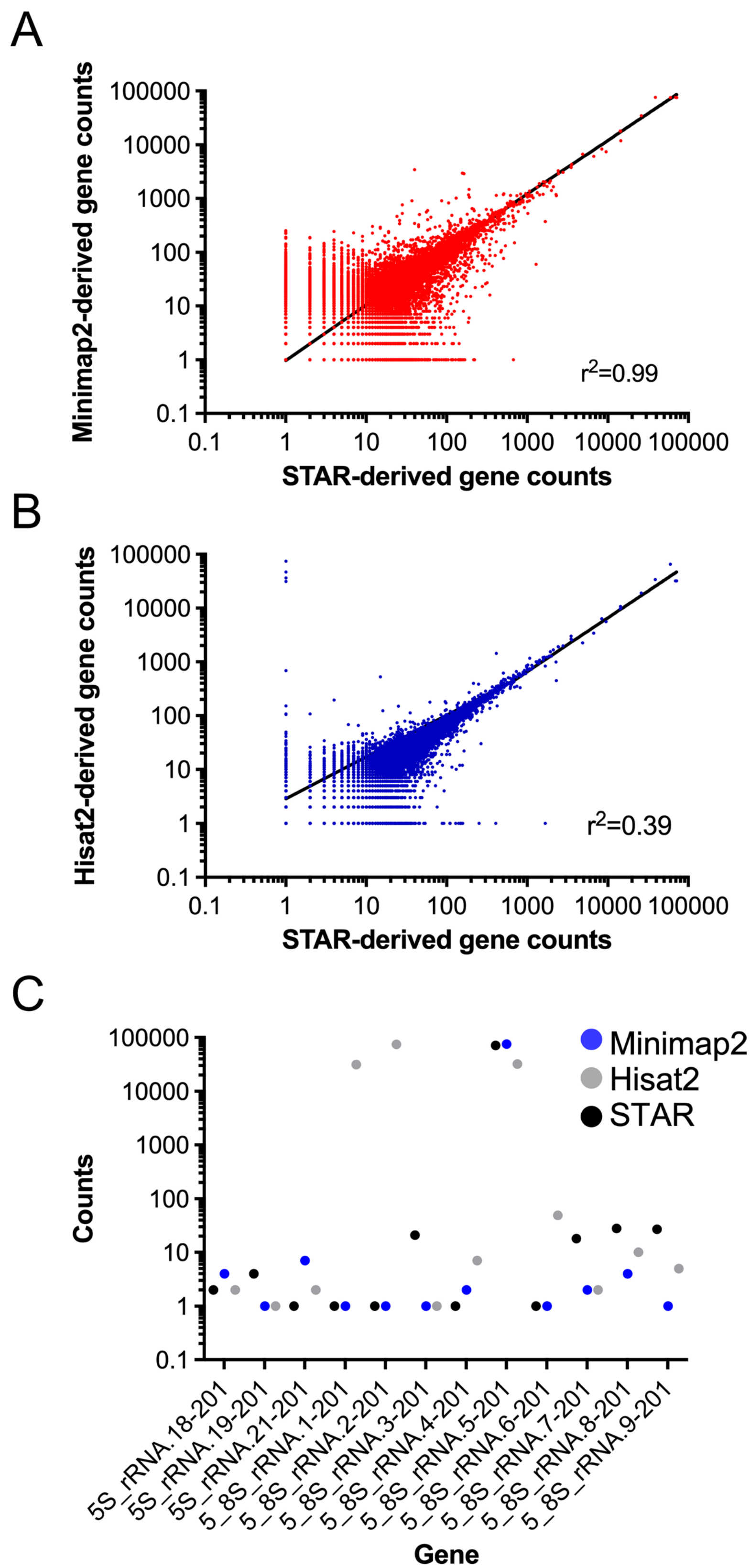


**Fig. 3** Comparison of RNA-Seq count data by different aligners. **A**) comparison of STAR and Hisat2 aligned data. **B**) Comparison of STAR and Minimap2 aligned RNAseq data. Raw count data were subjected to Linear fitting in Graphpad Prism v 6.0.**C**) Comparison of gene counts for various 5S-RNA species following alignment of data with STAR, Minimap2 and Hisat2

UMIs requires additional bioinformatics processing, to ensure the technical duplicates are properly identified and removed. Examples of UMI-based duplicate removal tools include UMI-tools [47], fgbio (https://github.com/fulcrumgenomics/fgbio), and Picard's MarkDuplicates WithMateCigar [48]. Some analysis platforms incorporate UMI demultiplexing (for example QiaSeq). It is worth noting that some single cell RNA-Seq studies have reported issues in the use of UMIs to determine gene expression levels [49].

*-Recommendation: Assessment of post alignment data can increase confidence of quality RNA library building, and interpretation accuracy. Where possible a small pilot can help assess library complexity and duplication issues prior to embarking on a full study.*

Following post alignment data assessment, one would then quantify the number of reads aligning to a given gene/known RNA. There are several quantification tools available, including HTSeq-count (https://htseq.readthedocs.io/en/master/htseqcount.html), featureCounts (https://subread.sourceforge.net/featureCounts.html), and RSEM (https://deweylab.github.io/RSEM/), and some alignment software can perform this automatically (such as STAR, which utilizes HTSeq to generate count data). These quantification tools primarily differ in how they handle multi-mapped reads. Non-weighted tools such as HTSeq-count and featureCounts either do not include any multi-mapped reads, include all multi-mapped reads, selects one location at random to include, or assigns a fraction for each multi-mapped read when generating read count data. Weighted tools such as RSEM utilize the maximum likelihood estimation method to assign multi-mapped reads to genes and/or transcripts when generating read count data.

RNA-seq raw count data should be normalized prior to downstream analysis to account for differences in sequencing depth, gene length, and RNA composition, allowing for the interpretation of gene expression levels. Common normalization methods include RPKM (Reads Per Kilobase of transcript per Million mapped reads), FPKM (Fragments Per Kilobase of transcript per Million mapped reads, used for paired end reads), and CPM (Counts Per Million). Each of these metrics enables comparison of gene counts between replicates of the same treatment group by adjusting for differences in sequencing depth, RPKM and FPKM also allow for comparison of gene counts between genes within a single sample since they also adjust for gene-length, but they are not suitable for differential expression analysis due to biases introduced during normalization. TPM (Transcripts Per Million) improves upon RPKM/FPKM by normalizing gene length first and then scaling by total transcript count, making TPM values more comparable across samples [50], however TPM is not suitable for differential expression analysis between different treatment groups since it does not adjust for RNA composition.

Many analysis tools now use raw gene counts matrices as the starting point for analysis and performs normalization as part of the analysis package (DESeq2/EdgeR/LIMMA) [51–53]. DESeq2 utilizes the median of ratios normalization method to adjust for sequencing depth and RNA composition, thereby enabling differential expression analysis between different treatment groups; however, an adjustment for gene length is needed for within sample comparisons. Although DESeq2 can optionally output FPKM values, its statistical testing is based on models using raw counts, not pre-normalized data. EdgeR utilizes the trimmed mean of M (TMM) normalization method to adjust for sequencing depth and RNA composition, and gene length corrected TMM can be applied to also adjust for gene length thereby enabling within sample comparisons. It is important to note that the type of normalization used can drastically impact the differentially expressed genes identified between groups, and the false discovery rate [54]. The choice of which software to use for analysis depends on the type of analyses and comparisons that need to be performed and cost [51]. One further consideration is that some tools perform better with smaller data sets. DESeq2 was developed for analyzing small n number RNA-Seq data sets and appear superior when sample sizes of less than 12 are used [55]. However, some benchmarking studies suggest an inflation of false positives when using DESeq2 with larger data sets [56]. Some GUI-based analysis packages can perform the entire pipeline of quality trimming and adapter removal to alignment to analysis. However, performing such analysis of a lot of data requires a dedicated server, or the use of cloud services. An intermediate between the two approaches is the Galaxy software suite [57], which hosts an impressive suite of analysis tools that the user can thread together to generate their own pipeline. The advantage of this approach is a quick learning experience, and Galaxy is often used as a first step for bioinformatics training [57].

RNA-seq does not provide absolute concentrations of an RNA or number of copies but rather allows the user to determine the relative abundance of an RNA within a sample of RNA. If an experimental condition results in a global change in RNA expression, this may affect subsequent downstream analysis. One solution to this is the use of spike in controls, such as the External RNA Controls Consortium (ERCC) [58–60]. Typically, these libraries of control sequences with known concentrations are added to an RNA sample (1:50-1000 wt/wt or 1–2% of reads) and sequenced alongside the sample [60]. Such standards do not appear to be affected by RNA sample complexity [60]. These standards consist of a range of RNAs at different concentrations, which allow for assessing the RNA-seq process (platform dynamic range and lower limit of detection), but also can be used for normalization of data. Multiple packages allow the normalization of data using spike-in controls, including LIMMA, DESeq2 and RUVSeq (58). When using spike-ins for data normalization, it is critical that extreme precision is used when adding spike-ins to samples and that the same spike-in mix is used for all samples. It is also important to ensure that the concentration of the RNAs in your samples are within the range of the spike-in concentrations. Thus, when using such approaches, it is recommended to assess this normalization method with other approaches to ensure the results are reliable as some limitations have been noted [58, 61, 62].

> *-The authors have used all three approaches for analysis of RNA-seq data. Each approach has merits and limitations. A lot depends on the individual's coding experience/comfort level (and library preparation experience when considering things like UMI-based deduplication and the use of spike-ins for normalization).*

### Accounting for batch effects

Although it is recommended to prepare and run all samples from the same RNA sequencing experiment together in the same run, that may not always be feasible due to the number of samples in the study, cost, and the nature of the experiment. If it is necessary to split the samples across different sequencing runs, samples should be split such that samples from each control and experimental group are included in each batch, as discussed previously. However, if equal distribution of samples from each group across batches is not feasible, the processed sequencing data should be evaluated for "batch effects", which is technical variation among samples [63, 64]. Analyses such as principal component analysis, multidimensional scaling, and hierarchical clustering dendrograms can be used to determine if samples cluster based on technical artifacts, such as variation in sample handling, preparation, or sequencing platforms, rather than by biological treatment groups. If batch effects are identified as a primary source of differences among samples, the technical batch effect(s) should be identified and accounted for prior to downstream analyses. There are several tools available to correct the technical artifacts that use different statistical approaches [65–69]. The type of tool and statistical approach that should be used for correction will depend on the nature of the technical variation and the distribution of the RNA sequencing data.

> *-The authors recommend testing several different approaches and evaluating how each correction approach affects the technical and biological differences in your study before making a final decision. For an example of this approach, visit our recent work that evaluated how different combinations of technical variable and batch effect correction algorithms impacted RNA sequencing data [70].*

## Other considerations

### Cost and sample size

The first question I am commonly asked is "*how much it will cost me?*", or "*how many samples can I get sequenced for a given budget?*" As with all scientific experiments it is useful to obtain a power calculation *apriori* to performing a study. There are online and R-based packages for calculating such statistics, to give confidence that given an effect size, you have sufficient experimental power to detect statistically meaningful differences [55, 71–73]. One key factor for consideration for RNA seq studies is dispersion, which refers to the variability in gene expression measurements across biological replicates. This variability can stem from biological differences (e.g., cell type heterogeneity or individual variation) or technical factors (e.g., differences in RNA quality or library preparation). Dispersion is typically gene- and tissue-specific and has a direct impact on statistical power. Higher dispersion reduces the confidence in point estimates of gene expression, which means larger sample sizes are needed to accurately estimate the true expression level of a gene and detect significant differences. Confidence intervals widen as dispersion increases, making it harder to detect true effects. Thus, power calculations must incorporate expected dispersion values for the system under study. Generally, genes with higher expression variability require more replicates to achieve acceptable power.

While dispersion reflects variability across biological replicates, reproducibility refers to the technical consistency of measurements across experiments or sequencing runs. It addresses the question: "If I repeat this experiment, will I get the same results?" High reproducibility ensures that technical variation is minimized, making biological differences easier to detect. Reproducibility can be improved through standardized protocols, automation, and laboratory experience. For example, consistent RNA extraction, library preparation, and sequencing procedures reduce technical noise. However, even with high reproducibility, some RNAs may exhibit inherently higher biological variability—especially in complex tissues—which still necessitates larger sample sizes.

Automation and experience can help reduce technical noise, but some RNAs are more variable than others, which could be a tissue-specific phenomenon. As a back of the envelope calculation, one is trying to detect

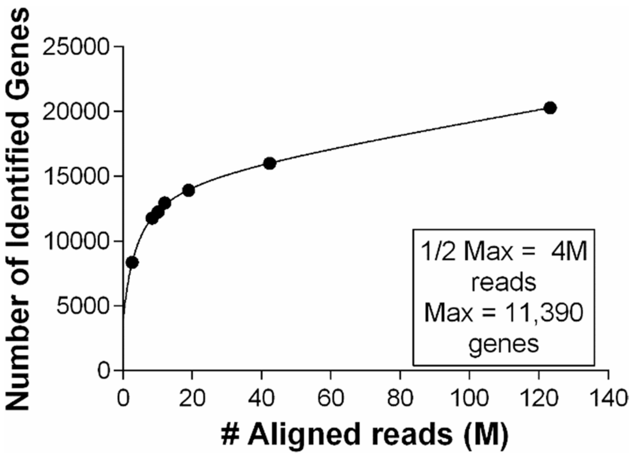


**Fig. 4** The number of identified RNAs in a given sample saturates with higher sequencing depth. Data from (6). Blood RNA seq data were aligned to the human hg19 reference and quantified using the Refseq annotation guide in Partek Genomics Studio. Various read depths were obtained by combining lanes of data from a SOLiD5500 run. Data were fitted to a mono-exponential curve in GraphPad Prism (v 6.0) to determine saturation and ½ max values

significant changes in a sampling of 10,000–20,000 RNAs, hence any power calculation should consider this correction for repeated measurement corrections. A recent study of murine RNA-Seq showed that a minimum of 6–7 and recommended 8–12 biological replicates is necessary to avoid a 50% false negative rate [74]. A second factor that will affect cost will be the choice of sequencing technology (number of samples which can be simultaneously assessed) and whether one is investigating technologies using short-read or long-read data. As noted above, long-read technologies typically have higher per-sample cost but may offer advantages in isoform resolution and transcript discovery.

*-Recommendation: Perform a power analysis prior to initiating an RNA-Seq study to understand the effect on power of the sample sizes one proposes to use.*

### Cost and sample depth

The depth of sequencing will greatly impact both cost and the type of analysis available for a given data type. For DNA seq a depth calculation is performed based on the number of reads you generate, multiplied by their size and divided by the size of the genome of study. Depending on the analysis 10-30x is deemed necessary for copy number variant detection and SNV analysis respectively (encode recommendations). RNA sequencing depth is typically stated by the number of generated reads. For gene expression typically 25–30 million aligned reads will saturate the number of genes detected in each sample (Fig. 4). Sequencing over this point, increases the number of genes identified in a linear manner, which may be due to noise (misalignment) or RNAs expressed at very low levels.

It should be noted that RNA-seq can be used for more than gene expression analysis, and such applications can affect the required read depth. For miRNA analysis a smaller total number of RNAs are identified hence depth is closer to 5–10 million reads. For structural analysis of short read data, a target of 100 million reads is recommended. For long read data 400,000–600,000 full length non chimeric reads per sample is recommended, but since barcoding is not available for direct RNA seq (yet) we typically generate around 20 million reads/sample using Promethion chips (recommendations for epitranscriptomics vary from). For discovery of novel noncoding RNAs, which may be expressed at very low copies/cell a sequencing depth of up to 300 M reads may be required. Abundance of the target is critical in these determinations. Finally, the number of RNAs in a given tissue may also depend on tissue complexity, more homogenous tissues may require slightly lower read depth, and more complex tissues will require greater read depth.

*-Recommendation: Factor into the cost of a study the depth of sequencing required to answer your biological question. Some abundant RNAs may require less sequencing, whereas discovery applications typically require a greater depth of sequencing.*

### RNA-Seq in non-model organisms

While our focus in this commentary is on human samples with well-annotated genomes, it is important to note that RNA-seq can also be effectively applied to non-human organisms lacking a complete reference genome or in situations where a complete reference genome is unavailable [30, 31]. In such cases, transcriptome assembly can be performed *de novo* [32], and the assembled isoforms used as a reference for alignment. Although this approach is essential for non-model organisms, reconstructed transcriptomes are also important in human studies to identify novel isoforms and better characterize alternative splicing events. This approach will be used to identify human -pan transcriptomes based on the human pan-genome project [75]. Additionally, mapping RNA-seq reads onto transcriptomes, even if incomplete, helps improve alignment accuracy by providing known exon-exon junction sequences [76, 77]. This guidance allows read spanning splice sites to be correctly placed, resolving junctions more reliably than genome-only alignment, which may miss or misalign spliced reads due to lack of annotated splice boundaries. Pseudoaligners such as Salmon and Kallisto offer rapid and efficient transcript quantification by avoiding traditional alignment and instead relying on k-mer-based methods to estimate transcript abundance. This dramatically reduces computational time and resource usage, making them highly suitable for large-scale studies. However, because these tools require a pre-existing transcriptome reference to function, they are limited to quantifying transcripts that are already known and annotated. As a result, their applicability is constrained in studies involving organisms without well-characterized transcriptomes or in contexts where novel transcript discovery is a key objective.

## Summary of commentary

The rapid development of RNA-Seq technologies and analysis tools has progressed at an amazing pace, leaving new researchers bewildered as to which path to take for a study. Here we have reviewed a few of the key technical considerations for the generation and analysis of RNA-Seq experimental data. While these are suggestions to help plan a study, ultimately access to technology and the experience of the investigator are key determinants of sequencing approach and depth of analysis.

**Acknowledgements**
Not applicable.

**Authors' contributions**
R.M., A.S. and R.V. wrote the main manuscript text, F.E. provided critical feedback and guidance. R.M. prepared the figures. All authors reviewed the manuscript.

**Funding**
This work was supported by NS116762, HG012334, HG013595, MD007602-33S1, NS112422, IOS-1956233,(PI Meller). Dr Meller is also supported by the CZI foundation. Additional institutional support at MSM provided in the form of grants from the NIH (MD000101 & MD007602). The sponsors of this study are public or nonprofit organizations that support science in general. They had no role in gathering, analyzing, or interpreting the data. The content of this publication does not necessarily reflect the views or polices of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government.

**Data availability**
No datasets were generated or analysed during the current study.

## Declarations

**Ethics approval and consent to participate**
Not Applicable.

**Consent for publication**
Not Applicable.

**Competing interests**
The authors declare no competing interests.

Received: 1 May 2025 / Accepted: 4 September 2025
Published online: 14 October 2025

### References

1. Shendure J. The beginning of the end for microarrays? Nat Methods. 2008;5:585–7.
2. Nazarov PV, Muller A, Kaoma T, Nicot N, Maximo C, Birembaut P, Tran NL, Dittmar G, Vallar L. RNA sequencing and transcriptome arrays analyses show opposing results for alternative splicing in patient derived samples. BMC Genomics. 2017;18:443.
3. Bullinger K, Dhakar M, Pearson A, Bumanglag A, Guven E, Verma R, Amini E, Sloviter RS, DeBruyne J, Simon RP, et al. Retrospective discrimination of PNES and epileptic seizure types using blood RNA signatures. J Neurol. 2025;272:128.
4. Gee BE, Pearson A, Buchanan-Perry I, Simon RP, Archer DR, Meller R. Whole blood transcriptome analysis in children with sickle cell anemia. Front Genet. 2021;12:737741.
5. Hardy JJ, Mooney SR, Pearson AN, McGuire D, Correa DJ, Simon RP, et al. Assessing the accuracy of blood RNA profiles to identify patients with post-concussion syndrome: a pilot study in a military patient population. PLoS ONE. 2017;12:e0183113.
6. Meller R, Pearson AN, Hardy JJ, Hall CL, McGuire D, Frankel MR, Simon RP. Blood transcriptome changes after stroke in an African American population. Ann Clin Transl Neurol. 2016;3:70–81.
7. Mina E, van Roon-Mom W, Hettne K, van Zwet E, Goeman J, Neri C, Mons PACtH, B. and, Roos M. Common disease signatures from gene expression analysis in huntington's disease human blood and brain. Orphanet J Rare Dis. 2016;11:97.
8. Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen LL, et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol. 2023;24:430–47.
9. Lee RC, Feinbaum RL, Ambros V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. Cell. 1993;75:843–54.
10. Wightman B, Ha I, Ruvkun G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell. 1993;75:855–62.
11. Michlewski G, Caceres JF. Post-transcriptional control of MiRNA biogenesis. RNA. 2019;25:1–16.
12. Iwakawa HO, Tomari Y. Life of RISC: formation, action, and degradation of RNA-induced silencing complex. Mol Cell. 2022;82:30–43.
13. Cui P, Lin Q, Ding F, Xin C, Gong W, Zhang L, Geng J, Zhang B, Yu X, Yang J, et al. A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. Genomics. 2010;96:259–65.
14. Huttenhofer A, Kiefmann M, Meier-Ewert S, O'Brien J, Lehrach H, Bachellerie JP, Brosius J. RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. Embo J. 2001;20:2943–53.
15. Ebbesen KK, Kjems J, Hansen TB. Circular rnas: identification, biogenesis and function. Biochim Biophys Acta. 2016;1859:163–8.
16. Gallego Romero I, Pai AA, Tung J, Gilad Y. RNA-seq: impact of RNA degradation on transcript quantification. BMC Biol. 2014;12:42.
17. Shen Y, Li R, Tian F, Chen Z, Lu N, Bai Y, et al. Impact of RNA integrity and blood sample storage conditions on the gene expression analysis. Onco Targets Ther. 2018;11:3573–81.
18. Ura H, Niida Y. Comparison of RNA-sequencing methods for degraded RNA. Int J Mol Sci. 2024. https://doi.org/10.3390/ijms25116143.
19. Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, et al. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nat Biotechnol. 2014. https://doi.org/10.1038/nbt.2972.
20. Schuierer S, Carbone W, Knehr J, Petitjean V, Fernandez A, Sultan M, Roma G. A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples. BMC Genomics. 2017;18:442.
21. Kobayashi T. Regulation of ribosomal RNA gene copy number and its role in modulating genome integrity and evolutionary adaptability in yeast. Cell Mol Life Sci. 2011;68:1395–403.
22. Shin H, Shannon CP, Fishbane N, Ruan J, Zhou M, Balshaw R, Wilson-McManus JE, Ng RT, McManus BM, Tebbutt SJ. Variation in RNA-Seq transcriptome profiles of peripheral whole blood from healthy individuals with and without globin depletion. PLoS ONE. 2014;9:e91041.
23. Saravia-Butler AM, Schisler JC, Taylor D, Beheshti A, Butler D, Meydan C, Foox J, Hernandez K, Mozsary C, Mason CE, et al. Host transcriptional responses in nasal swabs identify potential SARS-CoV-2 infection in PCR negative patients. iScience. 2022;25:105310.
24. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with tophat and cufflinks. Nat Protoc. 2012;7:562–78.

25. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, stringtie and ballgown. Nat Protoc. 2016;11:1650–67.
26. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. Stringtie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.
27. Ament IH, DeBruyne N, Wang F, Lin L. Long-read RNA sequencing: A transformative technology for exploring transcriptome complexity in human diseases. Mol Ther. 2025;33:883–94.
28. Chen Y, Davidson NM, Wan YK, Yao F, Su Y, Gamaarachchi H, Sim A, Patel H, Low HM, Hendra C, et al. A systematic benchmark of nanopore long-read RNA sequencing for transcript-level analysis in human cell lines. Nat Methods. 2025;22:801–12.
29. Clark MB, Wrzesinski T, Garcia AB, Hall NAL, Kleinman JE, Hyde T, Weinberger DR, Harrison PJ, Haerty W, Tunbridge EM. Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain. Mol Psychiatry. 2020;25:37–47.
30. Libro P, Chiocchio A, De Rysky E, Di Martino J, Bisconti R, Castrignano T, Canestrelli D. De Novo transcriptome assembly and annotation for gene discovery in salamandra salamandra at the larval stage. Sci Data. 2023;10:330.
31. Palomba M, Libro P, Di Martino J, Roca-Gerones X, Macali A, Castrignano T, Canestrelli D, Mattiucci S. De Novo transcriptome assembly of an Antarctic nematode for the study of thermal adaptation in marine parasites. Sci Data. 2023;10:720.
32. Raghavan V, Kraft L, Mesny F, Rigerte L. A simple guide to de novo transcriptome assembly and annotation. Brief Bioinform. 2022. https://doi.org/10.1093/bib/bbab563.
33. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 2019;20:278.
34. Amarasinghe SL, Ritchie ME, Gouil Q. long-read-tools.org: an interactive catalogue of analysis methods for long-read sequencing data. Gigascience. 2021. https://doi.org/10.1093/gigascience/giab003.
35. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21:30.
36. Wongsurawat T, Jenjaroenpun P, Wanchai V, Nookaew I. Native RNA or cDNA sequencing for transcriptomic analysis: a case study on *Saccharomyces cerevisiae*. Front Bioeng Biotechnol. 2022;10:842299.
37. Boulias K, Greer EL. Biological roles of adenine methylation in RNA. Nat Rev Genet. 2023;24:143–60.
38. Ewels P, Magnusson M, Lundin S, Kaller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32:3047–8.
39. Ward CM, To TH, Pederson SM. Ngsreports: a bioconductor package for managing FastQC reports and other NGS related log files. Bioinformatics. 2020;36:2587–8.
40. M M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17:10–2.
41. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30:2114–20.
42. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
43. Overbey EG, Saravia-Butler AM, Zhang Z, Rathi KS, Fogle H, da Silveira WA, Barker RJ, Bass JJ, Beheshti A, Berrios DC, et al. NASA genelab RNA-seq consensus pipeline: standardized processing of short-read RNA-seq data. iScience. 2021;24:102361.
44. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016;32:2103–10.
45. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.
46. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.
47. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. Genome Res. 2017;27:491–9.
48. "Picard Toolkit." 2019. Broad Institute. GithubRepository. https://broadinstitute.github.io/picard/(biotools:picard_tools; RRID:SCR_006525).
49. Tung PY, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, Gilad Y. Batch effects and the effective design of single-cell gene expression studies. Sci Rep. 2017;7:39921.
50. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods. 2008;5:621–8.
51. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
52. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
53. Chen Y, Chen L, Lun ATL, Baldoni PL, Smyth GK. EdgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. Nucleic Acids Res. 2025;53(2):gkaf018. https://doi.org/10.1093/nar/gkaf018
54. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-seq normalization methods from the perspective of their assumptions. Brief Bioinform. 2018;19:776–92.
55. Schurch NJ, Schofield P, Gierlinski M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. 2016;22:839–51.
56. Li Y, Ge X, Peng F, Li W, Li JJ. Exaggerated false positives by popular differential expression methods when analyzing human population samples. Genome Biol. 2022;23:79.
57. Galaxy C. The galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. Nucleic Acids Res. 2024;52:W83–94.
58. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014;32:896–902.
59. Pine PS, Munro SA, Parsons JR, McDaniel J, Lucas AB, Lozach J, Myers TG, Su Q, Jacobs-Helber SM, Salit M. Evaluation of the external RNA controls consortium (ERCC) reference material using a modified Latin square design. BMC Biotechnol. 2016;16:54.
60. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 2011;21:1543–51.
61. Consortium SM-I. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. Nat Biotechnol. 2014;32:903–14.
62. Kratz A, Carninci P. The devil in the details of RNA-seq. Nat Biotechnol. 2014;32:882–4.
63. Foox J, Tighe SW, Nicolet CM, Zook JM, Byrska-Bishop M, Clarke WE, et al. Performance assessment of DNA sequencing platforms in the ABRF next-generation sequencing study. Nat Biotechnol. 2021;39:1129–40.
64. Lai Polo SH, Saravia-Butler AM, Boyko V, Dinh MT, Chen YC, Fogle H, Reinsch SS, Ray S, Chakravarty K, Marcu O, et al. RNAseq analysis of rodent spaceflight experiments is confounded by sample collection techniques. iScience. 2020;23:101733.
65. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11:733–9.
66. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
67. Zhang Y, Jenkins DF, Manimaran S, Johnson WE. Alternative empirical Bayes models for adjusting for batch effects in genomic studies. BMC Bioinformatics. 2018;19:262.
68. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform. 2020;2:lqaa078.
69. Cuklina J, Pedrioli PGA, Aebersold R. Review of batch effects prevention, diagnostics, and correction approaches. Methods Mol Biol. 2020;2051:373–87.
70. Sanders LM, Chok H, Samson F, Acuna AU, Polo S-HL, Boyko V, et al. Batch effect correction methods for NASA genelab transcriptomic datasets. Front Astron Space Sci. 2023. https://doi.org/10.3389/fspas.2023.1200132.
71. Zhao S, Li CI, Guo Y, Sheng Q, Shyr Y. RnaSeqSampleSize: real data based sample size estimation for RNA sequencing. BMC Bioinformatics. 2018;19:191.
72. van Iterson M, van de Wiel MA, Boer JM, de Menezes RX. General power and sample size calculations for high-dimensional genomic data. Stat Appl Genet Mol Biol. 2013;12:449–67.
73. Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher JP. Calculating sample size estimates for RNA sequencing data. J Comput Biol. 2013;20:970–8.
74. Halasz G, Schmahl J, Negron N, Ni M, Atwal GS, Glass DJ. Optimizing murine sample sizes for RNA-seq studies revealed from large-scale comparative analysis. bioRxiv. 2024:2024.2007.2008.602525.

75. Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, Popejoy AB, Asri M, Carson C, Chaisson MJP, et al. The human pangenome project: a global resource to map genomic diversity. Nature. 2022;604:437–46.
76. Zhao S. Assessment of the impact of using a reference transcriptome in mapping short RNA-Seq reads. PLoS ONE. 2014;9:e101374.
77. Pham MT, Milevskiy MJG, Visvader JE, Chen Y. Incorporating exon-exon junction reads enhances differential splicing detection. BMC Bioinformatics. 2025;26:193.

## Publisher's Note