Chapter 12: Guidelines and important considerations for 'omics-level studies

Abstract

RNA sequencing (RNA-seq) is a high-throughput sequencing technique used to analyze and quantify the entire population or subsets of cellular RNA. This chapter presents experimental protocols and data analysis methods for poly-adenylated RNA-seq, with a focus on research questions related to quantification of gene and isoform expression, comparison between groups, and studies of the genetic regulation of gene and isoform expression. A comprehensive discussion of study design considerations introduces important concepts and presents real-life situations to show researchers how to obtain well balanced and calibrated datasets. These best practices apply beyond RNA-seq and also provide a solid foundation for other high-throughput sequencing studies.

The advent of high-throughput sequencing has made it possible to conduct unbiased, genome-wide measurements of a large range of molecular phenotypes. While these techniques are increasingly easy to implement in all labs familiar with standard molecular biology experimentation, the complexity of these datasets motivates many important considerations when undertaking ‘omics-level studies. In this chapter, we will review several experimental and analytical best practices, with a focus on RNA sequencing studies for gene expression and mRNA splicing.

RNA sequencing (RNA-seq) refers to high-throughput sequencing of either the entire population or subsets of cellular RNA. This chapter focuses on the most common implementation of RNA-seq: complementary DNA (cDNA) sequencing after capture and reverse-transcription of polyadenylated messenger RNA (mRNA). This technique allows for the quantification of steady-state levels of mature mRNA molecules (fully transcribed, capped and polyadenylated), but often is not powered to analyze lowly expressed, transient, or quickly degraded mRNA species. RNA-seq data can be used not only for the quantification of gene expression levels, but also for identifying the composition of expressed genes or mRNA isoforms, quantification of specific mRNA isoforms, and evaluation of the relative inclusion of specific exons or splice sites in mRNA molecules. Thus, RNA-seq provides insight into molecular processes ranging from transcription, mRNA splicing and processing, and mRNA degradation.

Before high-throughput sequencing approaches, mRNA profiling was performed with microarrays spotted with a set number of probes in the 3’ untranslated regions (UTRs) of genes or alternatively spliced exons. Following hybridization of cellular cDNA, fluorescence intensity was used to quantify gene expression or exon usage. These methods were dependent on prior knowledge of gene and isoform structure, prohibiting discovery of novel transcribed regions, isoforms, or alternatively spliced exons. Furthermore, if the annotation used to create the probe set was incorrect, measurements could bias biological insight into the gene expression or splicing. Finally, microarray approaches only allowed for limited genome-wide analyses, since the number of probes was limited by the microarray scaffold. RNA-seq approaches overcome these difficulties and allow for unbiased sequencing of cDNA that allow for simultaneous mRNA discovery, annotation, and quantification.

Overview of the RNA-sequencing experimental protocol

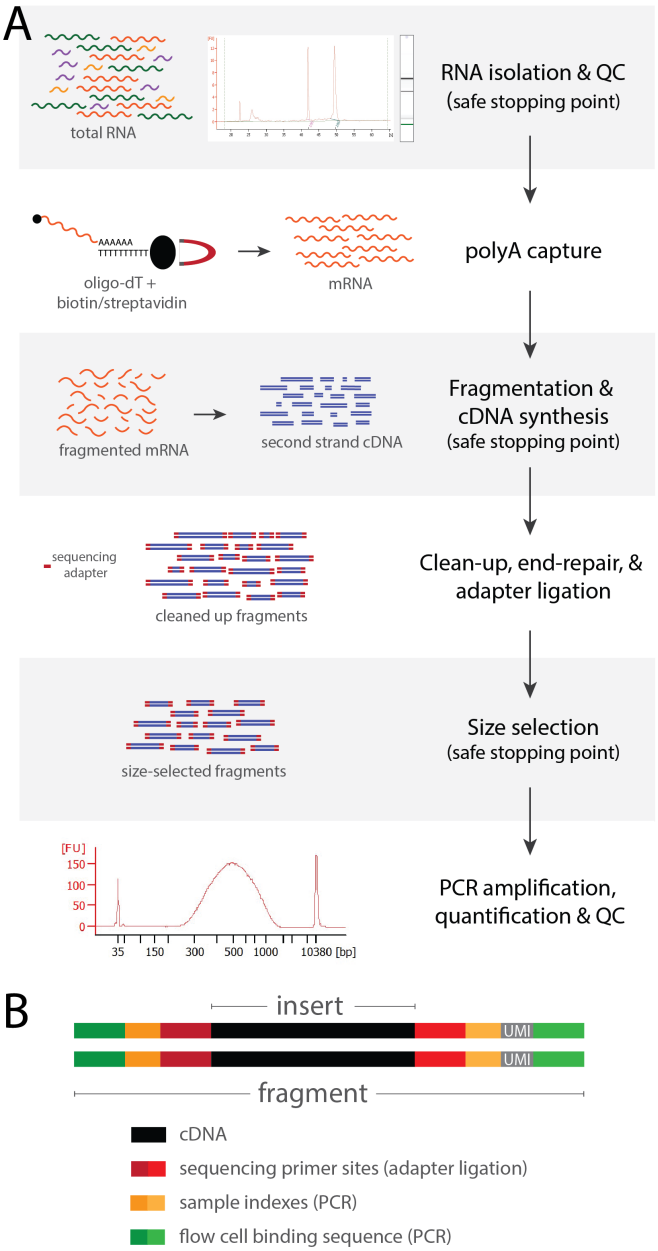

Figure1: RNA-seq experimental protocol. (**A)** Protocol steps with examples of Agilent Bioanalyzer results for a high quality RNA sample (RIN>9, *top*) and for a final library (*bottom*). The average fragment size of the library is 500bp, which is ideal for splicing and ASE analysis, in addition to gene expression quantification and differential analyses. This protocol is usually executed over 3 days, safe stop points indicated. (**B)** Final library construct design, including dual indexes (*yellow*) and UMIs (*grey*).

RNA-seq library preparation protocols share several steps with other high-throughput sequencing (HTS) protocols, but also pose unique challenges due to the delicate nature of the starting material. Here we will review the major steps of the protocol to prepare libraries from polyadenylated mRNA, with a special focus on the steps that can be modified/adjusted for specific applications **(Figure 1A)**. One of the challenges for researchers that are getting started on RNA-sequencing is the length and complexity of the experimental protocol. However, aside from a few individual steps, the majority of the steps require basic molecular biology techniques.

It is crucial that RNase free consumables and reagents are used at all times. This includes standard reagents like water, which should be molecular grade and nuclease-free for all steps. Whenever possible, RNA-seq protocols should be performed in a dedicated laboratory space with partition between pre- and post-PCR steps to decrease the risk of contamination.

RNA isolation is the first step in the protocol. Most RNA-seq applications use total RNA as starting material, followed by either mRNA capture or rRNA depletion. rRNA depletion is preferred when the study aims to profile the entire RNA population, including small RNAsand other RNA molecules that may not be poly-adenylated. When limited material is available, it is possible to capture mRNA directly from cell lysate, thus skipping the total RNA isolation step and increasing RNA yield. This is usually a safe stopping point in the protocol and isolated RNA can be preserved at -80C°.

The RNA is then fragmented and primed for cDNA synthesis. cDNA synthesis is performed with random primers, however particular applications may require using targeted primers. At this step it is also possible to add Unique Molecular Identifiers (UMI) to allow for direct counting of RNA molecules **(Figure 1B)**. Following first and second strand cDNA synthesis, the library preparation protocol can be stopped and the cDNA stored at -20C for up to 1 week.

The cDNA is cleaned with SPRI beads, followed by end repair and adapter ligation. This is a critical step as the adapter design depends on the library preparation kit and sequencing strategy of choice. During this step, or the subsequent PCR step, short oligonucleotides (indexes) can be added to each library to identify them when pooling multiple libraries in a single sequencing run. Ideally, dual indexing should be used to maximize the probability that reads are correctly assigned to their sample of origin despite sequencing errors in the indexes **(Figure 1B)**. This is particularly important when a large number of samples are pooled together in the same sequencing run and for certain sequencing platforms.

At this point it is necessary to enrich fragments within a specific size range in the library. This step serves two purposes: (1) Removing potential adapter dimers that formed during ligation because of imbalances in the cDNA/adapter ratio, which may occur if cDNA synthesis had sub-optimal efficiency and (2) Removing large cDNA fragments, which could create steric interference on the sequencer during cluster generation. The correct ratio of beads to DNA ratio is used to efficiently remove very short fragments that are preferentially sequenced and could drastically reduce the sequencing yield. Deciding the upper limit for size selection depends on the scientific question. Short reads provide adequate information for most applications, however splicing and allele-specific analyses require longer reads (i.e. 300 cycles) to increase the probability of sampling the splice junction or polymorphic site, respectively. Special consideration should be paid to both the size-selection and bead clean-up steps, as the techniques used are specific to HTS protocols. Beads are used to bind the nucleic acid (i.e. cDNA) and, depending on the beads:nucleic acid ratio, they can also be used to select a specific range of fragment lengths. Generally, size selection is most efficient for large fragment sizes (right side of the fragment size distribution) as opposed to removing short fragments (left side of the fragment size distribution). It is advisable to practice bead clean-up and size selection prior to performing the first RNA-seq experiment.

The adapter-ligated DNA is ready to undergo PCR enrichment, followed by quality control of the library and determination of optimal loading concentration. The quality control step aims to confirm that the library produced is within the concentration range recommended for sequencing. It is advisable to use a capillary electrophoresis approach to QC the library, rather than an agarose gel or spectrophotometer to visualize and quantify the library, respectively. A library with low concentration may be indicative of low quality/concentration RNA, inefficient cDNA synthesis, issues with bead clean up steps or failed adapter ligation. Additionally, this step confirms that the prepared library is of the expected size and that adapter dimers were efficiently removed during size selection.

Study design considerations - definitions

The key to a successful 'omics study is integration of experimental and data analysis considerations at an early study design stage. Many experimental details or protocol decisions that seem purely technical in nature are actually deeply connected to the biological question being asked. Ignoring these considerations may often lead to false positive results. Designing a well powered study begins with access to high quality samples and is strongly dependent on the technical steps that constitute the experimental execution of the project. Here, we will focus on key experimental aspects that bear a

significant weight in the overall success of a study and discuss how to integrate them in the data analysis considerations, which will be further expanded in the next section.

RNA quality. In some cases a project starts with the investigator having full control over the sample collection. In this best case scenario, it is of critical importance to optimize the sample preparation so that the RNA integrity is preserved. However, it is more often the case that the researcher only has access to previously collected samples, thus having limited control over the RNA quality. RNA integrity is crucial to the success of an RNA-sequencing experiment. It is important to note that RNA-sequencing is a less forgiving experiment than quantitative Real-Time PCR.

It is standard in the field to report RNA quality using an RNA Integrity Number (RIN), which varies on a scale of 1-10 (10 being best and 1 being worse). While the RIN of RNA isolated from cell cultures or model organisms in the laboratory is usually very high (9 or greater), the RIN can be quite low for RNA isolated from other sources. Nevertheless, RNA-sequencing of post-mortem samples in humans has yielded high quality libraries with minimal RIN values of 6 (\relax GTEx Consortium 2013). When RIN varies greatly across samples, it is advisable to include it as a technical variable/confounder in statistical models for RNA-seq analysis (Gallego Romero et al. 2014). RNA quality also affects library complexity and data quality. A library with low complexity results in a high proportion of reads from PCR duplicates, which are identical copies of independent cDNA fragments but do not contribute independent information to the overall measure of gene expression. To limit the occurrence of PCR duplicates, RNA-seq protocols minimize the number of cycles during the PCR enrichment step. However when low concentration or low quality RNA is used as input, even a low number of PCR cycles may result in a high proportion of PCR duplicates or to library preparation failure. Another feature of RNA-seq data that is influenced by RNA quality is the proportion of reads mapping to exons. When this number is low, it may indicate genomic DNA contamination, RNA degradation and/or low efficiency in the poly-A RNA capture or rRNA depletion. Thus, the proportion of reads mapping to exons is another important technical variable to consider when modeling gene expression during data analysis.

Confounders. Technical variables for RNA quality are an example of experimental confounders that may complicate the analysis of RNA-seq data. The most well known experimental confounders are batch effects, which is an umbrella term for any grouping of experimental samples that may result in features of the data that are shared within the group and do not have direct biological relevance for the research question. Batch effects can be introduced at any step along the experimental protocol, starting with sample collection, RNA isolation, library preparation and sequencing. Working in batches is inevitable in functional genomics experiments, where the number of samples and experimental conditions considered is usually large. Therefore, while batch effects are unavoidable, it is critical to limit batch effects with careful study design, to record confounders for further consideration during data analysis, and to avoid batch effects that are confounded with the research question. In particular, when designing a study, batches should be balanced for all variables relevant to the research question **(Figure 2)**. When relevant variables are too numerous, they can be randomized across batches. For example, in a study aimed at identifying gene expression differences across multiple cell types, it may be difficult to have each cell type represented in a batch because of different growth rates. In this case, the cell types included in each batch could be randomized to avoid batches that contain exclusively or mostly one cell type.

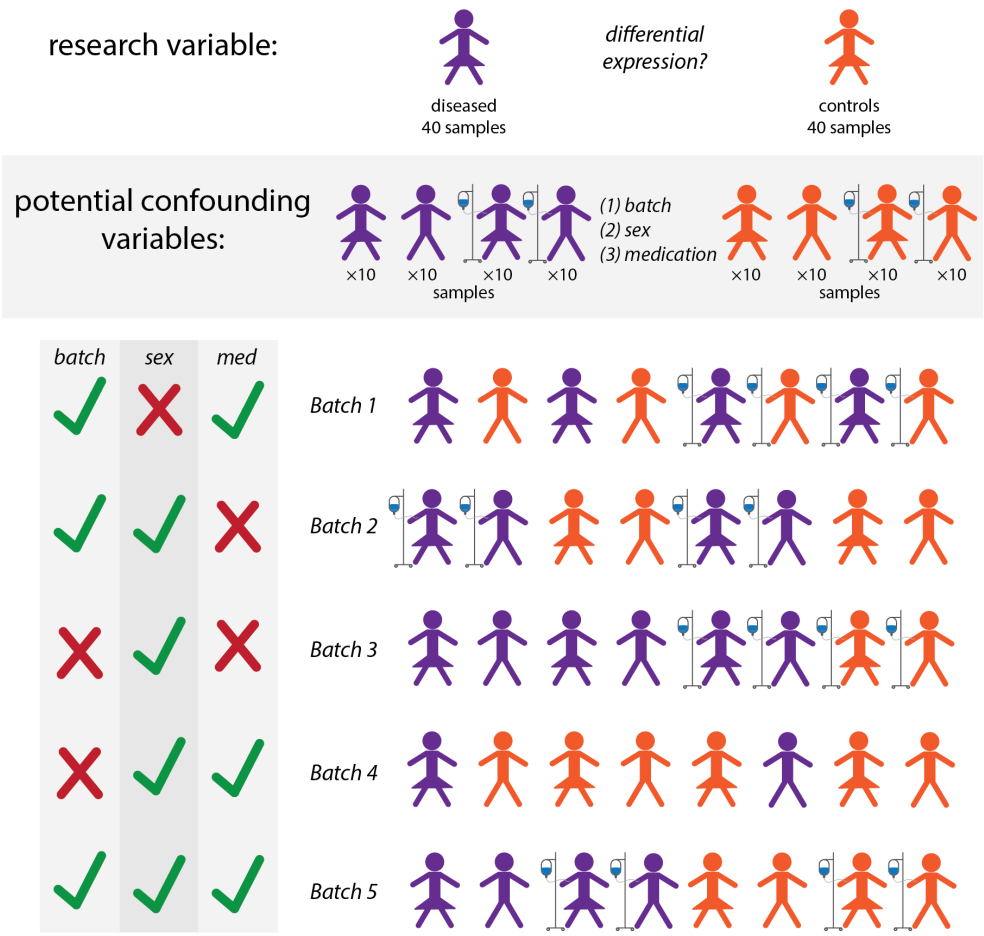

Figure 2: Study design and confounders. Example of a study design to identify differential gene expression between diseased and control patients. The sample size is 40 for each group (balanced sample size). Potential confounding variables are the patient's sex and medication status. The bottom panel presents different batch designs. For each batch design, the left column indicates if the study design is balanced (*green check*) or if the study variable (disease status) is confounded with any of the other variables (sex, medication status, batch) (*red cross*). In a well designed study, the variable of interest (disease status) is not confounded with any additional variable.

Finally it is crucial to carefully define the research question(s) prior to beginning sample collection. This will ensure that sample groupings in batches do not confound group contrasts. The example in Figure 2 presents a hypothetical study aimed at identifying gene expression differences between cases and controls for a pathological condition, thus disease status is the variable of interest. Let's imagine that after the data are collected the investigator decides to use this dataset to also study the effect of medication on gene expression. If the batches are designed only considering disease status as the variable of interest, it is likely that some batches only contain medicated subjects, while others are over-represented in unmedicated subjects **(Figure 2, Batch 4)**. Therefore it would be impossible to investigate differences in gene expression due to medication. However, if the batches are balanced also by medication status, they can be used to investigate this additional research question. It is also possible that changes in gene expression associated with disease are also correlated to differences due to medication status, so the best practice would be to control for this variable regardless of the additional research question.

Replicates. Replicability is one of the key concepts in science. A rigorous finding is based on results that have been replicated in independent experiments or sample groups. Because of the complexity of genomics experiments, replication may take several forms and occur at multiple steps along the experimental process. Depending on which segment of an experiment is replicated, replication may have caveats or yield conservative results. Researchers across scientific fields have not reached a consensus on the definition of technical and biological replicates. Intuitively, these two terms are based on the sources of variability that may contribute to a lack of reproducibility. If the source of variability is biological, repeating the experimental step to replicate the sampling of this biological variability would be considered a biological replicate. If the source of variability is technical - does not originate from a biological process but is inherent to the experimental procedure - replicating the relevant experimental

step would result in a technical replication. Some examples may help illustrate these points. In the example above, experiments performed on samples from different individual donors from the same disease status group are considered biological replicates. If multiple aliquots of the same cell pellet are used to independently isolate RNA and prepare RNA-seq libraries, these would be technical replicates of the RNA isolation and library preparation steps. However, other types of replication experiments are possible and their definition becomes more controversial. For example, for some researchers, independent growth of cell cultures from the same donor would be considered biological replicates of the culturing process, while for other researchers they are technical replicates of the experimental procedure to grow cells. Ideally, one would perform replicates of different experimental steps to identify the most relevant sources of variation and focus on those for all subsequent experiments. However this strategy may not always be feasible and researchers often prefer full biological replicates, when possible.

Sequencing depth. When designing an 'omics experiment, the experimenter must also consider the final characteristics and amount of data needed to answer specific biological questions. For high-throughput sequencing experiments, it is important to think about the number of short reads needed to estimate biological parameters at genome-wide scale. This is referred to as *sequencing depth*, defined as the number of reads sequenced from a library. Importantly, the depth is determined by the experimenter, through two main choices. First, the total read output differs between sequencing instruments. For Illumina short-read sequencing, the number of reads sequenced on a single run can range between 4 million on an iSeq 100 instrument to 2.6 billion on a NovaSeq 6000. Second, users can determine the proportion of reads from a lane or flowcell that are allocated for each library by multiplexing libraries. *Multiplexing* refers to the practice of mixing multiple different libraries to sequence them together. This is achieved by using unique library-specific indexes during library preparation (usually during the PCR step as described above) **(Figure 1B)**, which are then used to disambiguate libraries while mapping (see below). If multiplexing, it is critical not to combine two libraries with the same indexes and most ideal to use a set of indexes that all have at least two nucleotide differences from another index. Indexes are generally 6 or 8-mers. While there is no limit to the number of libraries that can be mixed together, higher multiplexing reduces the number of reads obtained from each library and can present challenges for the downstream computational separation of reads from individual libraries.

The ideal sequencing depth of a library is determined by the anticipated complexity of a library and genomic coverage. Library *complexity* refers to the number of unique molecular fragments within a library, driven by the amount of starting material and the target molecular enrichment (ie. mRNA, small RNAs, protein binding peaks, etc), see above. Libraries prepared with very low amounts of material or targeting low frequency genomic regions are likely to have lower complexity, leading to saturation of unique biological information with lower numbers of reads. Standard RNA-seq libraries are usually extremely complex, owing to the large diversity and dynamic range of mRNAs in most cells. However, there are exceptions – for instance, RNA-seq libraries from red blood cells generally have low complexity since red blood cells are enriched for hemoglobin mRNA (Uellendahl-Werth et al. 2020; Harrington et al. 2020). Genomic *coverage* refers to the number and distribution of reads across all expected genomic regions. Higher sequencing depth usually results in higher coverage – specifically, a more even distribution of reads across all regions (not just highly expressed genes) and enough reads at each region to provide power to make biological conclusions. For DNA sequencing experiments,

coverage is usually a primary consideration, where Nx coverage refers to N reads overlapping every base or region on average.

Decisions about how deeply to sequence a library are usually driven by a balance between statistical power and experimental cost for replicate samples. Too few reads would lead to lower coverage and thus prohibit robust and reproducible quantification of genes or isoforms (Sims et al. 2014). This would particularly affect the quantification of low to medium expressed genes and result in low confidence for exon-specific analyses in these genes, which rely on rare splice junction reads. On the other hand, it is expensive to achieve higher sequencing depths for each library and potentially not needed. For highly complex RNA-seq libraries, studies have found that gene expression can be robustly estimated using 10-15 million fragments (10-15 million single end reads or 20-30 million paired end reads) and isoform or splicing analyses can be conducted with 30-50 million fragments on average. A similarly high sequencing depth (40-100 million fragments or more) is required also for allele-specific expression analysis, in order to confidently calculate allelic ratios. If preliminary analyses indicate that more reads are necessary, it is always possible to re-sequence the same library to obtain more reads, but it is necessary to ensure that potential confounds are taken into account before combining data from two runs (see above).

Read length and type. Two important sequencing parameters are the length of the reads and the number of reads sequenced from each independent fragment. Read length is determined by the number of sequencing cycles run on the sequencer and can range between 30 – 300 nucleotides on an Illumina machine. Note that specific sequencing kits must be used for higher cycle numbers, with different instruments being able to sustain longer sequencing runs (ie. MiSeq will sequence 300 nt reads). Fragments in the library can either be sequenced from only one end (single end read) or from both ends (paired end read). When choosing read length and single vs paired end reads, it is important to consider the average *insert size* of the library, defined as the length of the cDNA fragment between the two adapters (fragment length – total adapter length; **Figure 1B**). Read lengths longer than the insert size will result in sequencing into the adapter, which can be avoided with shorter reads. Similarly, if the insert size is shorter than the sum of the paired end reads, the read pair will be providing redundant information, artificially increasing the coverage at the overlapping region. If aiming for a specific read length for single or paired end reads, then it is necessary to prepare the library with an appropriate fragment length (see above). After sequencing, software tools like Picard (Toolkit 2019) can be used to infer the insert length from the data to determine whether adapter trimming might be necessary or paired ends reads are likely to be overlapping.

Decisions about read length and pairing influence the robustness of the downstream analyses. First, these parameters influence the ability to confidently map or align reads to the genome. Longer reads are more likely to map to unique positions in the genome. Similarly, paired end reads can help to position reads since both reads can be used to disambiguate between similar regions. For instance, if one read maps to a homologous or repeat region, the second read may be used to properly position the entire fragment if mapped to a unique region. Second, longer reads and paired end reads aid with isoform quantification and splicing analyses. Longer reads are more likely to overlap an informative splice junction, providing direct evidence of a splicing event. Analogously, long reads are more likely to overlap one or multiple heterozygous site(s) for allele-specific analysis purposes. Paired end reads can be used to quantify splicing of an exon even when both reads are within flanking exons by inferring

whether the pair of reads could have arisen from an unspliced vs spliced transcript given the average insert length of the library.

Analysis of RNA-seq data

Similar to other 'omics applications, data generation is just the first step. A complex and careful data analysis plan is necessary to pre-process and QC the data and to test the hypothesis or identify patterns of biological interest. Importantly, data analysis should not be seen as independent from data generation, as they are two highly interconnected steps of the overall RNA-seq experiment. Ideally the experimentalist and data analyst are either the same person or work very closely in designing the experiment, recording potential confounders, and interpreting the data. This section presents key concepts and approaches to RNA-seq data analysis, many of which are also relevant for other 'omics applications.

Quality control. Before initiating any biological analyses, it is important to ensure that high-throughput sequencing data is of high quality. As described above, there are several experimental and technical considerations that could influence data quality and bias gene expression levels. Experimental considerations like unhealthy or dying cells and poor RNA quality must be assessed prior to sequencing a library, since they are impossible to assess in the final data but can have a large impact on biological interpretation. In contrast, sample preparation issues resulting in DNA, rRNA, or adapter contamination in the final library should be avoided but can be assessed in the final library as described below. When such sample preparation issues are identified, reads can either be filtered or libraries can be re-prepared after DNase digestion, further rRNA removal or polyA selection, and selection for larger fragment sizes to correct the issues listed above, respectively. It is customary to run several quality control checks on raw sequencing data before proceeding with mapping and quantifying mRNA levels. These quality control checks include examining the distribution of read quality scores (for confidence in base calls), evaluating overrepresented sequences (to assess library complexity), adapter reads, adapter contaminated reads (at the ends), and read quality near the ends of reads. There are many packages (ie. fastQC, RNA-seQC (Andrews and Others 2017; Graubert et al. 2021)) that streamline the implementation of these checks and provide visual analyses for quick evaluation of data quality.

Mapping of reads. The first step in any high-throughput sequencing analysis workflow is to “map” reads to genomic coordinates or align reads to reference genome or transcriptome sequences. While there are dozens of mapping software that have been written for short read sequencing data, each has specific properties that are important to consider based on the biological question. Here, we focus on two considerations that are specifically crucial for RNA-seq analyses. First, when mapping RNA-seq data, it is important to use a splicing-aware mapper (i.e. TopHat (Trapnell et al. 2009), STAR (Dobin et al. 2013), HiSAT2 (Kim et al. 2019)) to map splice junction reads, which contain large genomic gaps that cannot be handled by standard genome mappers **(Figure 3A)**. Though it is possible to use a non-splicing-aware mapper to map RNA-seq reads to a transcriptome reference sequence instead of the genome, it is more advisable to map to a genome reference sequence. Mapping to the genome with a

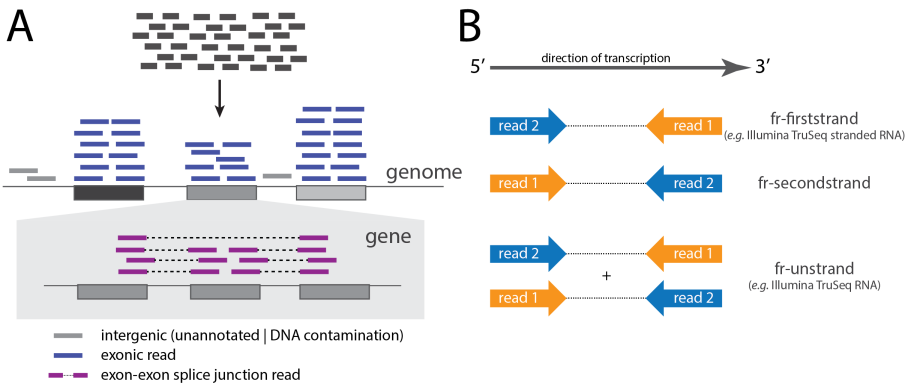

Figure 3: Mapping and Strandedness. (A) RNA-seq reads include intergenic (DNA contamination), exonic and exon-exon splice junction reads. A splicing-aware software allows proper mapping of exon-exon splice junction reads and quantification of DNA contamination. **(B)** Libraries can be prepared using stranded or unstranded protocols. The details of the protocol are important to infer correct strand information from mapped data or input into downstream quantification or read counting algorithms.

splicing aware mapper allows for quantification of novel transcribed regions or splice junctions after scaffolding on known splicing events and allows for the quantification of DNA contamination (reads mapping to intergenic or intronic regions) **(Figure 3A)**. However, when sequencing RNA from a species that has no reference genome or a poorly annotated one, it may be advisable to use the RNA-seq data to assemble a reference transcriptome, scaffolded on the genome or transcriptome of a closely related species. Second, for genetic analyses f (e.g. allele-specific expression), it is important to avoid mapping artefacts that lead to a biased allelic representation among the mapped reads. Since mapping software rely on a reference genome, reads are more likely to be mapped properly when they exactly match the reference alleles. This can be alleviated using a haplotype or genotype aware mapper (i.e. HISAT2), which accounts for both allelic options during mapping. Furthermore, this problem can be exacerbated when there are experimentally driven substitutions in the data, such as in SLAM-seq experiments where uridines labeled with 4sU appear as T > C substitutions in the sequencing data (Herzog et al. 2017). In these cases, it is useful to downweigh known substitutions using a mapper such as NextGenMap (Sedlazeck et al. 2013) (not splicing aware) or HISAT3n (Zhang et al. 2021) (splicing aware). Lastly, following mapping, it is worth looking at raw data on a genome viewer (such as iGV (Robinson et al. 2011)) to identify any issues and evaluate your expectations.

Note that read alignment software will assign reads to the strand that directly matches the read sequence. However, the assigned strand may not reflect the strand from which the parent transcript was transcribed. If the library was not prepared with steps to maintain strand information, reads have an equal probability of mapping to either strand. If a stranded library preparation was used, the transcriptional strand can be identified. Libraries can be prepared using one of two strandedness protocols, which result in the first read matching either the sense or antisense strand **(Figure 3B)**. Thus, it is important to identify the experimental specifics to infer correct strand information from mapped data or input into downstream quantification or read counting algorithms.

Quantifying gene and isoform expression. When using RNA-seq data to quantify mRNA expression levels and compare the expression levels across genes or samples, it is crucial to account for two parameters. First, read counts must be normalized by the total number of reads sequenced. If sample 1 has been sequenced to a depth of 50 million reads and sample 2 only has a total of 10 million reads, each gene is likely to have 5 times more reads in sample 1 than sample 2 even if the abundance of all mRNAs is equivalent between samples. Second, it is necessary to account for the total length of a gene or isoform. If gene 1 is longer than gene 2, there is a higher probability of sequencing reads from gene 1 independent of gene expression levels. Thus, mRNA expression levels are generally computed with

one of two metrics that accounts for these confounding variables: Reads per kilobase per million (RPKM) and Transcripts per million (TPM). *RPKM:* a straightforward metric that simply divides read counts by the total number of reads sequenced for the library (in millions) and then by the length of the transcribed region from which reads are counted (in kilobases) (Mortazavi et al. 2008). A variation of this is fragments per kilobase per million (FPKM), which is applied to paired end reads and uses the count of read pairs rather than individual reads, since paired reads are not statistically independent from each other. *TPM:* a similar metric, where read counts are first divided by length and then by the total number of reads sequenced (Li and Dewey 2011). While this seems like an inconsequential mathematical change, the difference in normalization order greatly improves the ability to compare relative gene expression levels across samples. The second normalization by the total number of million reads forces the sum of TPMs within each sample to be 10^6 (not true for RPKM or FPKM metrics), which standardizes the proportional levels across samples. Importantly, the need to normalize by library-specific parameters makes it only possible to quantify relative expression levels (rather than absolute) for standard RNA-seq libraries. For these reasons, TPM has become the preferred metric to quantify gene expression from RNA-seq data.

Either of these quantification metrics can be calculated in several ways. To estimate RPKM or TPM by hand, the researcher first needs to count reads for each exon or gene feature (using software such as ht-seq (Anders et al. 2015) or featureCounts (Liao et al. 2014)) and then use the formulas above to calculate the desired metric. These read counts, either using only uniquely mapping or including reads that map to multiple locations, can also be used as input for differential expression analyses, for which statistical models are run on raw rather than normalized counts (see below). A more robust method of quantification involves using a statistical maximum likelihood model to probabilistically assign multi-mapping reads across the multiple locations to which they map. This approach is implemented in software like RSEM, which uses an expectation-maximization algorithm to calculate maximum likelihood abundance estimates from mapped reads scaffolded on known transcriptome annotations (Li and Dewey 2011). This approach will output: (1) adjusted read counts that accounts for the expected read count based on both uniquely and multi-mapping reads and (2) TPM values using these adjusted read counts. Finally, a more extreme implementation of the expectation-maximization algorithm is used in reference-free pseudoalignment-based isoform and gene quantification approaches (ie. Kallisto (Bray et al. 2016) or Salmon (Patro et al. 2017)), which do not rely on an initial alignment step but match kmers within reads to compatible transcripts to obtain maximum likelihood isoform or gene abundance estimates. By circumventing the mapping step, this reference-free approach alleviates issues arising from mapping biases of individual regions and instead focuses solely on quantification of known isoform sequences. Pseudoalignment quantification approaches are extremely fast and result in TPM levels for isoforms, however they do not provide a list of mapped regions with their genomic coordinates. Since both maximum likelihood approaches quantify gene expression using a full cohort of reads rather than only those mapping to unique regions, they are more likely to provide robust abundance estimates. Furthermore, these approaches enable isoform level quantification, which is harder to do with read counts alone where exon features can be shared across isoforms. It is important to note however, that isoform level quantification always relies on known annotations and specifically known junctions between exons. When these annotations are incorrect or incomplete, they can bias isoform TPMs. However, it is always possible to add the isoform TPMs from a single gene to get a gene-level TPM, which is more robust to annotation biases since it relies only on knowledge of exonic regions rather than specific splicing patterns.

Quantifying absolute mRNA abundance. All quantification methods within samples highlighted above are inherently relative measurements, since they must account for differences in sequencing depth and other library-specific properties. Specifically, metrics such as TPMs are designed to describe the relative proportion of mRNA that arises from a given gene – a gene with TPM = 1 represents 1 millionth of the total mRNA population, while TPM = 1000 represents a thousandth of the total population. Thus, an increase in TPM between samples supports an increase in the relative proportion of mRNA represented by that gene but does not allow any conclusions to be drawn about the absolute amount of mRNA produced from that gene. To perform absolute quantification, libraries must be designed to account for differences in cell count and total RNA yield between samples, the loss of material during RNA extraction and library prep, and any biases that occur during sequencing. The most popular method to do this is to spike a pool of exogenous RNA into a sample at the beginning of RNA extraction or library preparation. This "spike-in" RNA can be from a different species (preferably far enough diverged from the species of interest to allow abundant sequence dissimilarity) or a population of synthetic RNA designed to account for different transcript lengths and sequence compositions (ie. the controls designed by the External RNA Controls Consortium (ERCC) (External RNA controls — the joint ini...)). Furthermore, if performing a polyA selection, the exogenous RNA must have polyA tails. A small amount of exogenous RNA is spiked in relative to the number of cells (or a similarly relevant parameter like weight of tissue) used to extract RNA from the sample of interest, which allows calibration for differences in total RNA yield and thus absolute RNA abundance between samples (Lovén et al. 2012). This addition should happen as early as possible in sample preparation (i.e. cell lysis) so that the spiked-in RNA goes through the entire sample preparation process simultaneously with the sample of interest. Sequencing reads from a library with spiked-in RNA should be mapped to a genome where the spike-in sequences or genome of the spike-in species have been combined with the genome of interest as distinct and labeled chromosomes. To obtain an abundance estimate that accounts for absolute differences in RNA abundance, TPMs from the samples of interest can then be normalized by the proportion of reads that map to the spike-in samples.

Comparing gene expression between groups. A common application of RNA-seq is to compare expression levels between two groups of samples and identify differentially expressed genes. This general framework applies to several research questions, for example identifying gene expression changes induced by a treatment *in vitro* or *in vivo* or comparing expression in treated and untreated samples. Another common scenario is the comparison of gene expression between cases and control groups to identify genes that are differentially expressed in a disease state. Finally, gene expression levels are also commonly compared across developmental stages and tissue or cell types. Generally, the researcher will identify a dichotomous variable of interest, which will be used to partition the samples into two groups. Because of the confounders discussed above, the variable of interest must be defined before collecting the RNA samples and performing the experiment, thus ensuring that any potential batch effect is not confounded with the variable of interest. Read length is not a critical factor for these studies, as even the shortest reads (75 cycles split over pair-end reads on the Illumina NextSeq500) are sufficient for accurate gene expression quantification and comparison between groups. Many studies use a sequencing depth of 20-40 million paired-end reads per sample, however studies have demonstrated that even <10 million reads per sample are sufficient to determine whether the variable of interest is associated with gene expression differences between the two groups (e.g. (Moyerbrailean

et al. 2016)). If the researcher is interested in a detailed characterization of gene expression changes between two groups, greater sequencing depth is necessary to capture smaller effects.

Prior to performing formal statistical tests for differential gene expression, a few data processing/visualization approaches are helpful to identify major axes of variation in the dataset, determine if the data present any unexpected structure (due for example to an unmeasured confounder) and confirm which confounding variables should be included in the statistical model. Principal component analysis of the gene counts and a scatter plot of the first few (1-4) PCs can be used to visualize if the samples form unexpected clusters. Correlation analysis of the PCs with the variable of interest and potential confounders can also be used to quantitatively determine the main axes of variation in the data. Pairwise correlation between all samples, followed by visualization, is a complementary method to visualize the data structure and identify unexpected clustering of the samples. This can be further investigated through hierarchical clustering on the correlation matrix.

To identify genes differentially expressed between the two groups, a linear model is used to model expression of each gene as a function of the variable of interest, of measured confounders and error. Common packages such as EdgeR (Robinson et al. 2010) and DESEQ2 (Love et al. 2014) implement a fixed effect linear model, which does not allow random effects between individuals to be modeled. This is generally acceptable for most study designs; however if multiple measurements of the same individual are included (i.e.in time-series experiments or experimental replicates), a random effect linear model or a mixed effect linear model are more appropriate, as they will account for the correlation structure that exists in the data due to multiple measurements from the same individual. LIMMA (Ritchie et al. 2015) is one of the most common packages used to analyze RNA-seq data using a random effect linear model. One key difference between these two classes of methods is also the data pre-processing. Methods such as DESEQ can be applied directly on the count matrix and use a negative binomial model to estimate the overdispersion parameters in the data, thus not requiring an additional normalization step. The LIMMA package includes a few different normalization approaches that are performed prior to testing for differential gene expression. In both cases, the output will be a measure of the expression change for each gene, usually indicated as average log2(fold change) and the statistical significance (p-value). To account for the large number of tests conducted, it is important to perform multiple test correction. Two common multiple test correction approaches are the Benjamini-Hochberg p-adjusted to control for the false discovery rate (Benjamini and Hochberg 1995) and Storey's q-value (Storey 2003). The final list of differentially expressed genes will be defined at a certain false discovery rate (q-value or p-adjusted) and potentially also after setting a threshold on the fold change in gene expression if the researcher is interested in filtering effects based on the magnitude of the change. To ensure that the tests are well calibrated, it is common to create a qq-plot to compare the observed p-value distribution to the expected uniform distribution. Volcano plots are also used to visualize the relationship between fold change and p-value and set the above mentioned thresholds.

While gene expression comparisons between two discrete groups is a common framework for many research questions, sometimes the variable of interest may not be dichotomous (e.g. presence/absence of treatment) but varies continuously in the entire sample. Examples of continuous variables are time/age (e.g. in time-series experiments), dosage (e.g. in dose-response experiments), serum analytes levels and many others. In these cases, there are two options. First, the variable can be encoded as continuous and a linear model is used to test for an association between the variable of

interest and gene expression over a continuous scale. Alternatively, the variable can be discretized by creating arbitrary breaks that may reflect experimental or biologically meaningful categories (e.g. low, intermediate, high level of serum analyte). In both cases, the underlying assumption is a linear relationship between gene expression and the variable of interest. However this assumption may not always be true. For example, in a dose-response curve, the expression of a gene may increase initially and then plateau or even decrease, depending on the underlying regulatory mechanism and potential feedback loops. A simple but limited approach to this problem is to reduce these types of datasets to multiple pairwise comparisons and compare each data point to the previous in the series or to the baseline. While this is still an ongoing area of statistical development, clustering approaches have also been used effectively to identify patterns of gene expression in series of measurements.

Studying the genetic determinants of gene expression. An application of RNA-seq that has become widely popular in the last decade is expression quantitative trait locus (eQTL) mapping **(Figure 4)**. With the decrease in sequencing costs and the ability to perform RNA-seq on a large number of samples, researchers have paired RNA-seq data with genotype data to identify associations between genotype at a locus and expression of a nearby (cis-eQTL) or distant (trans-eQTL) gene. Here we will discuss eQTL mapping in general with a focus on cis-eQTLs since those are the easiest to identify, but many of the considerations presented are also applicable to trans-eQTL mapping. Furthermore, the general QTL mapping approach, can also be used to map loci associated with other molecular phenotypes, this can be achieved by combining genotype data with other types of 'omics data (e.g. splicing QTLs, chromatin accessibility QTLs, etc). Generating RNA-seq data for eQTL mapping presents unique challenges compared to the other applications discussed here. First, these studies need a much larger sample size. There is good consensus in the scientific community, supported by both power calculations and large scale studies (e.g.GTEx), that a sample size of at least 70 individuals is sufficient to detect large genetic effects on gene expression for a large number of genes. Though eQTL mapping studies have also used thousands of samples, they are still quite smaller than the hundreds of thousands of samples used for genome-wide association studies. Working in batches for eQTL mapping is inevitable, therefore all the study design considerations discussed above apply. Furthermore, if the goal is to compare genetic effects on gene expression across conditions (response-eQTL mapping, reQTL), the different conditions should be processed in parallel within the same batch. Sequencing depths

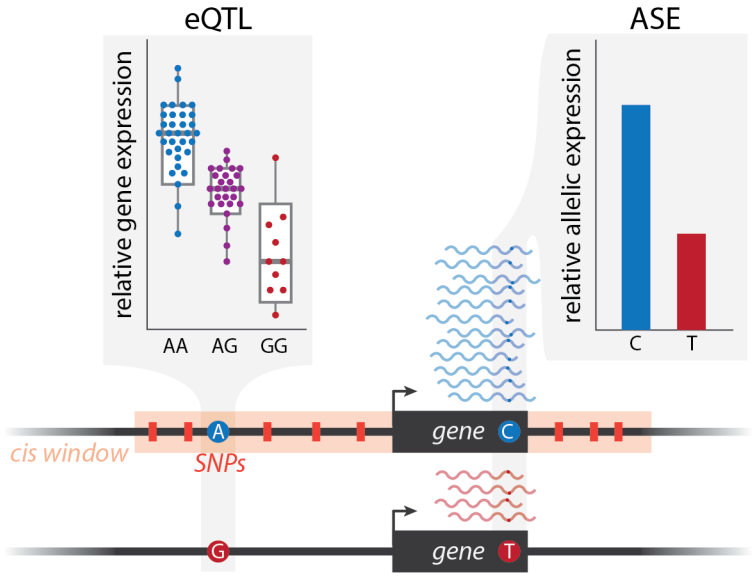

Figure 4: Methods to characterize genetic regulation of gene expression. In the eQTL example (*left*), each dot of the boxplot represents an individual. The median gene expression is indicated by the line in the box and is calculated across individuals within each genotype class. A significant association is present between genotype at the non-coding SNP and gene expression across individuals, with higher expression associated with increasing number of copies of the A allele. The ASE example (*right*) shows the distribution of reads at a coding heterozygous site within one individual sample. A significant allelic imbalance is present between the C and T alleles, with higher expression of the C allele. In the absence of additional information, the eQTL has a higher probability of being the true causal variant compared to the coding SNP.

and read lengths appropriate for the comparisons between groups are also appropriate for eQTL mapping.

Because genetic effects on gene expression are generally small, unmeasured and measured confounders have a large impact in this application of RNA-seq. The effects of confounders are removed via an iterative procedure that maximizes the number of significant eQTLs by progressively removing a larger number of confounders. Confounders can be quantified through several methods, including principal component analysis, surrogate variable analysis (Leek and Storey 2007) and peer factor analysis (Stegle et al. 2010). The statistical test for eQTL mapping is implemented in a few different packages, including FastQTL (Ongen et al. 2016) and MatrixQTL (Shabalin 2012). Essentially these are different flavors of a linear model where gene expression is the dependent variable, while the independent variables are the genotype dosage (0, 1 or 2 number of copies of the minor allele), and the confounders. The first step is to define the cis-regulatory region, which is usually arbitrary defined between 100Kb and 1Mb. The RNA-seq data can be normalized using the same procedure as in the LIMMA pre-processing (commonly Voom normalization). Association between genotype and expression is tested for each gene/SNP pair within the cis-regulatory region. Because the tests are not independent, due to linkage disequilibrium between SNPs in the same cis-regulatory region, the p-value distribution is often inflated. Permutations are used to correct the p-value distribution, followed by Storey's q-value method for multiple test correction.

Identifying response eQTLs, which are genetic loci associated with a difference in gene expression levels upon stimulation or cellular perturbation, is more challenging and there is limited consensus on the methodological approach. The traditional approach uses the same statistical framework used for eQTL mapping but with expression change as the dependent variable. However this approach has limited power and can only capture the most extreme cases, when the association between genotype and gene expression is only present in one condition or has opposite sign in the two conditions considered. In most cases, the genetic effect on gene expression undergoes more subtle changes across conditions, thus more sophisticated methods are needed. Bayesian methods have been successfully applied to this problem, and have demonstrated to have the potential to identify condition/context-specific genetic effects across several groups and beyond the two groups comparison e.g. (Urbut et al. 2019).

An alternative and complementary approach to QTL mapping is allele-specific analysis **(Figure 4)**. As for QTL mapping approaches, allele-specific analysis can be performed on different types of 'omics data. Allele-specific expression (ASE) analysis is used to identify genes with regulatory variants from RNA-seq data. ASE analysis is performed only on genes that contain heterozygous variants and assumes that the true causal site is a variant in the regulatory region, which is also heterozygous in the sample under consideration. Under the null hypothesis of absence of regulatory variation at a gene locus, the two alleles at the heterozygous coding site should be represented by a similar number of reads in the RNA-seq data. ASE is defined as a departure from this 50:50 allelic ratio, and is formally tested with a beta-binomial model that accounts for over-dispersion in the allelic read counts. It should be noted that in rare cases, ASE may result from technical or biological processes that are independent of gene regulatory variants (e.g. imprinting). ASE analysis can be performed in a single individual sample, which is an advantage over eQTL mapping. Furthermore, trans-effects are fully controlled

because allelic effects are compared within the same individual. However, detection of ASE for a gene is limited by the number of polymorphic sites in the coding region and by their heterozygosity.

Analysis of alternative splicing events. While some analyses of mRNA splicing patterns can be conducted by using isoform-level TPM values, it is often more precise to delve into individual exon-level changes to understand alternative splicing changes. There are several types of exon-level changes that are often studied, which can be broken down into three main categories: (1) alternative transcription start sites, including alternative first exons, (2) alternative splice sites, including alternative 3' or 5' splice sites, cassette or skipped exons, mutually exclusive exons, and retained introns, and (3) alternative polyadenylation sites, including alternative last exons and tandem 3' untranslated regions **(Figure 5)**. While the first and the third are usually quantified using read coverage in the relevant exons, analyzing alternative splice site usage relies on combined exonic and junction read coverage. Popular metrics to quantify alternative splicing include percent spliced in (PSI as in (Katz et al. 2010)) or percent alternative usage (PAU as in (Ha et al. 2018)), which broadly calculate the percentage of reads that support a defined inclusion isoform (i.e. inclusion of a cassette exon) relative to the total informative reads in the region.

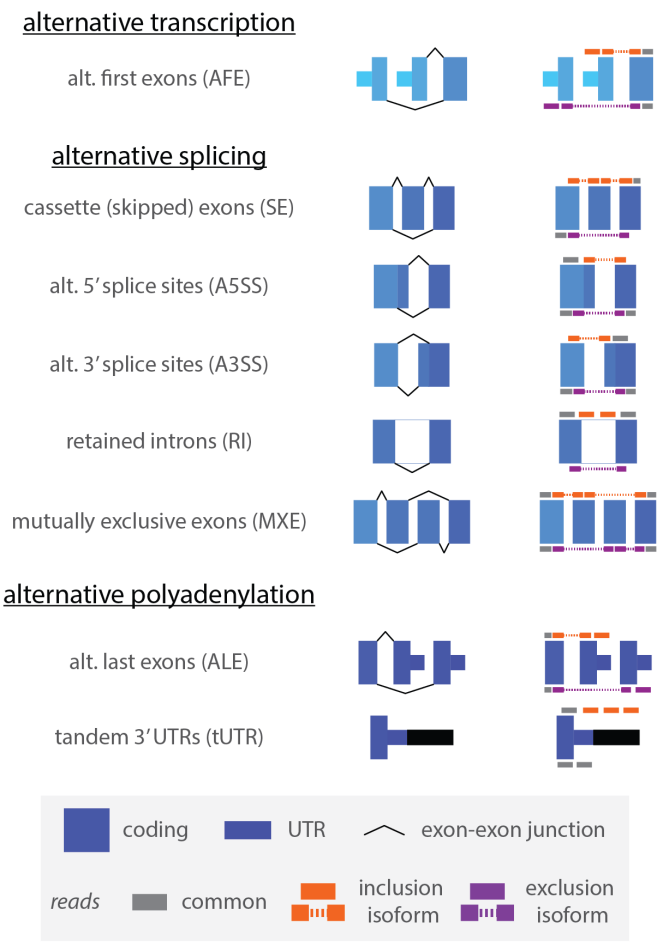

Figure 5: Alternative splicing and other exon-level changes. Examples of alternative transcription, splicing and polyadenylation events that can be quantified in RNA-seq data. The left column indicates the splicing isoforms and the right column shows the reads that are informative for quantifying the usage of inclusion (*top, orange*) or exclusion (*bottom, purple*) isoforms.

Quantification software generally takes one of two approaches: (1) local read analysis or (2) isoform-anchored local event analysis. In the first approach, exonic and junction reads specific to a particular event are used to assign an PSI value, where the definition of alternative events and inclusion vs. exclusion isoforms is based on known annotations (Katz et al. 2010; Shen et al. 2014). Since this method relies heavily on junction reads, many software packages have options to initially discover novel events in a dataset and incorporate these into downstream quantification (Vaquero-Garcia et al. 2016; Li et al. 2016). Crucially, all quantification occurs only at a local level, agnostic to what happens upstream or downstream of the specific event of interest and thus avoiding any biases caused by long-range isoform annotation. However, since the approach gains power through a high density of informative reads within a small region, many events do not have sufficient read density to allow for quantification. The second approach uses isoform level TPM quantification (as described above) to infer a quantification of local event usage by aggregating information from isoforms that include vs exclude a given exon or site (Alamancos et al. 2014). Since the definition of isoforms is necessary, this approach often precludes denovo identification and quantification of events. Additionally, the reliance on long-range annotations could bias exon-specific

quantification if isoform structures are incorrect or incomplete. However, isoform-anchored local event analysis is often much quicker and easier to implement, making it a good option for a first pass analysis. Furthermore, since isoform level TPM quantification uses reads across the entire isoform, these approaches often have more power to quantify alternative splicing in lower expressed genes. Finally, these approaches are often better suited for measuring terminal exon usage of either alternative first or last exons since those event types may not have informative junction reads (Ha et al. 2018; Goering et al. 2020). Regardless of the method used to quantify PSI or PAU values, these values can be used for splicing QTL analyses with the same considerations and statistical approaches described above.

Summary

In this chapter, we summarize the experimental and computational considerations for high-throughput RNA-seq libraries. We focus primarily on the most common poly(A) RNA-seq libraries, which are enriched for mature mRNA molecules. However, there are many other types of RNA-seq libraries that either enrich for different RNA species (i.e. total RNA population, small RNAs) or target RNA at different stages of their lifecycle (i.e. chromatin associated RNA, nuclear or cytoplasmic RNA, ribosome associated RNA). Here, we walk through the specific protocol and important considerations for poly(A) library preparation. Importantly, we go through several best practices for the study design of these experiments, including measuring the quality of RNA prior to starting libraries, potential biological and technical confounders, replicates necessary for statistical analyses, and sequencing parameters. Designing a well controlled and well calibrated study is important for any high-throughput experiment, so most of these considerations apply to all high-throughput sequencing studies. Furthermore, they will likely all continue to be important regardless of future experimental or computational developments in the field. In our last section, we outline steps for analyzing RNA-seq datasets, with a focus on initial QC, gene expression measurements, differential expression and splicing, and using these data to study the influences of genetic variation on mRNA levels and compositions.

Finally, it is important to note that there is constant development of new high-throughput sequencing protocols to sequence different subsets of RNA, with many new technologies being published and widely adopted each year. Similarly, statistical analysis methods, software, and study design standards are an area of active research and new techniques may correct previous biases or improve biological interpretability of results. Thus, it is crucial to survey the literature for new developments prior to undertaking an RNA-seq experiment to ensure conformation with the latest best practices. Our hope is that the considerations presented in this chapter can provide a useful guide to identify important and relevant aspects when keeping up to date in the field.