1. Intro on RNA-seq (500) - keep it short because may overlap w/ chapter 11 - Athma
   1. what it is, what can we learn from this data, advantage over microarrays, different types of RNA-seq
   2. Focus on Poly-A RNA-seq

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 leads to 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 (see box). The most common implementation of RNA-seq is complementary DNA (cDNA) sequencing after capture and reverse-transcription of polyadenylated messenger RNA (mRNA), which we will focus on in this chapter. 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 for the quantification of gene expression levels, but also for identifying the composition of expressed genes or mRNA isoforms, quantify specific mRNA isoforms, and evaluate 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 done with microarrays spotted with a small number of set probes in the 3’ untranslated regions (UTRs) of genes or alternatively spliced exons. Following hybridization of cellular cDNA, florescence intensity was used to quantify gene expression or exon usage. These methods depended on knowledge of gene and isoform structure, prohibiting discovery of new transcribed regions, isoforms, or alternatively spliced exons. Furthermore, if the annotation used to create the probe set was incorrect, measurements could bias biological interpretation of 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.

1. Box on other types of RNA-seq
   1. What they are good for, what's different in terms of protocol/samples prep, what's different analytically, reference to actual protocol
   2. Total RNA-seq
   3. miRNA-seq (cross link w/ dedicated chapter)
   4. MPRA/Targeted RNA-seq

1. Protocol (500) - Fra - done
   1. Example protocol (appendix)
   2. Major steps - w/ focus on the actionable ones
   3. Protocol Workflow figure - where can you stop?

* Similar for other types of RNA-seq
* Size selection considerations
* How to starting with small amounts

1. Study design considerations - definitions (1000)
   1. RNA quality - Fra -done
   2. Confounders - Fra-done
      1. Batch effects - sample, sequencing, RNA-extraction, library prep -
      2. Sample collection
      3. Variation between groups or samples
   3. Replicates - technical vs biological - Fra-done
   4. Sequencing depth/multiplexing -> question – Athma

**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 number of reads differs between sequencing machines. For Illumina short-read sequencing, the number of reads sequenced on a single run can range between XX on an I-seq instrument to XX on a NovaSeq. 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), 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). Libraries started 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 mostly comprised of hemoglobin mRNA. 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 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. Too few reads would lead to lower coverage and thus prohibit robust and reproducible quantification of genes or isoforms. 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 reads and isoform or splicing analyses can be conducted with 30-50 million reads on average. 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).

* 1. Read length & single/paired end – Athma

The next important sequencing parameter is 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 machine 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 machines 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). If the Read lengths longer than the insert size will result in adapter sequences on the reads, so it might be better to choose shorter read lengths. Similarly, if the insert size is shorter than the sum of the paired end reads, the read pair will be providing redundant information. 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 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 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. 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.

1. Questions and study design (4000-5000).
   1. mRNA quantifications (1000) - Athma
      1. Analysis metrics/methods - basic principles (then examples of software)

**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 doing 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. If identified, reads can either be filtered or libraries can be re-reprepared 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 include checking for the distribution of read quality scores (for confidence in base calls), over-represented 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, others?) 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 with known genomic or transcriptomic sequences. While there are dozens of mapping software that have been written for short read sequencing data, each has different properties that are important to consider based on your biological question. While there are many considerations, here we will highlight two 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, STAR) to be able to map splice junction reads, which contain large genomic gaps that cannot be handled by standard genome mappers. 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 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). Second, for genetic analyses like questions of 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 when sample genotypes are known by using a haplotype or genotype aware mapper (i.e. HISAT2), which accounts for both allelic positions during mapping. Furthermore, this problem can be exacerbated when there is 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. In these cases, it is useful to downweigh known substitutions using a mapper such as NextGenMapper (not splicing aware) or HISAT3n (splicing aware).

* Strandedness – types of strandedness
  + 1. Getting TPMs, difference between TPM & RPKM, etc

**Quantifying gene and isoform expression.** When using RNA-seq data to quantify mRNA expression levels, it is crucial to account for two crucial parameters to compare the expression levels across genes or samples. 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 similarly is a higher probability of seeing more 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):* 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). 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. *Transcripts per million (TPM):* a similar metric, where read counts are first divided by length and then by the total number of reads sequenced. 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 106 (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 or featureCounts) 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. 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 implemented in pseudoalignment-based isoform and gene quantification approaches (ie. Kallisto or Salmon), which do not rely on an initial alignment step but matches kmers within reads to compatible transcripts to obtain maximum likelihood isoform or gene abundance estimates. By circumventing the mapping step, this approach alleviates issues arising from mapping biases of individual regions and instead focuses solely on quantification of known isoform sequences. Though they do not result in files with mapped regions, pseudoalignment quantification approaches are extremely fast and result in TPM levels for isoforms. 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 exons regions rather than specific splicing patterns.

* + 1. Spike-ins to normalize: relative vs. absolute quantification

**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 RNA can be RNA 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 ERCC RNA). 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 (ora similarly relevant parameter like weight of tissue) used to extra RNA from the sample of interest, which allows calibration for differences in total RNA yield and thus absolute RNA abundance between samples. 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.

* 1. Comparing groups (1000)  - Fra
     1. Differential gene expression
     2. Replicates
     3. Dichotomous vs continuous variables
  2. eQTL mapping (1000) – Fra
  3. Splicing/RNA processing (1000) - Athma

**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. 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) or percent alternative usage (PAU), 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.

Quantification software generally take one of two approaches: (1) local read analysis or (2) isoform-anchored local event analysis. The first approach is to use exonic and junction reads specific to a particular event to assign an PSI value, where the definition of alternative events and inclusion vs. exclusion isoforms is based on known annotations. 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. 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. Since the definition of isoforms is necessary, this approach often precludes denovo identification and quantification of events and the reliance on long-range annotations could bias exon-specific quantification if isoform structures are incorrect or incomplete. However, these approaches are often much quicker and easier to implement, making them 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.