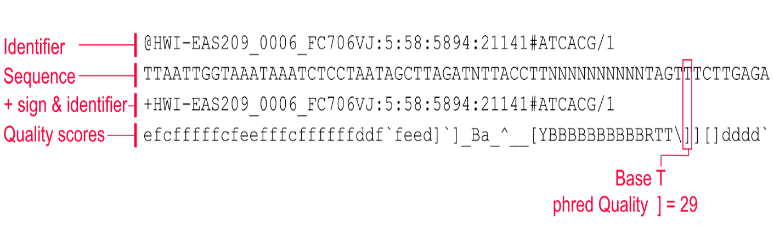
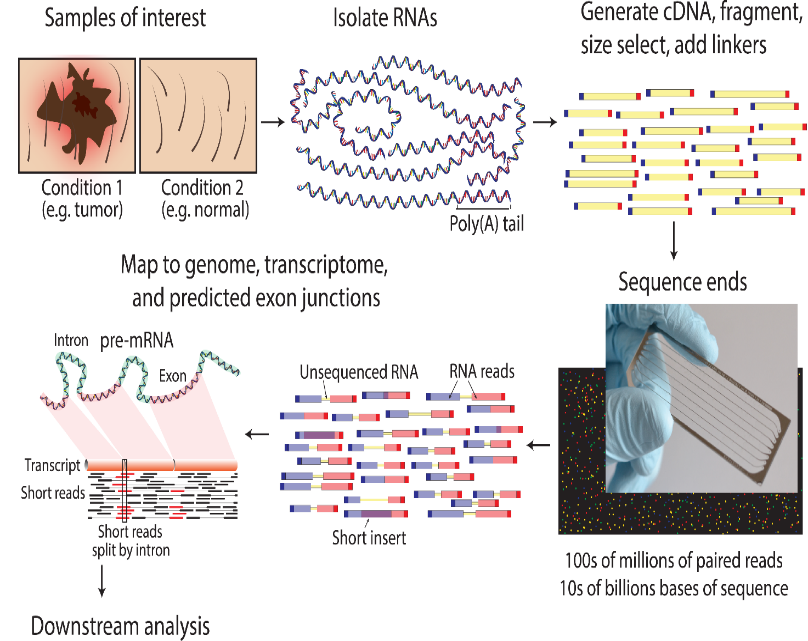
mRNA sequencing

**Understanding mRNA sequencing in broad strokes:**



FASTQ files

1. **Sample preparation:** start with a sample containing RNA. In my case PHHs.
2. **Isolate mRNA:** The total RNA extracted from cells or tissues contains various types of RNA molecules, including ribosomal RNA (rRNA), transfer RNA (tRNA), small nuclear RNA (snRNA), and messenger RNA (mRNA), among others. In most mRNA sequencing experiments, the primary interest is in the mRNA fraction because it represents the genes being actively expressed in the cells.
3. **Reverse transcription:** Convert the mRNA into complementary DNA (cDNA using reverse transcriptase.
4. **Library preparation:** The cDNA (or mRNA) is fragmented, and adapters are added to each fragment. These adapters are crucial for the sequencing process.
5. **Sequencing**: The prepared library is loaded onto a sequencing machine (here a Illumina NovaSeq 6000). Each fragment is sequenced, producing millions of short sequences.
6. **Data analysis**:Bioinformatics tools are used to align these short sequences to a reference. This alignment process helps determine.

**Sample preparation**

Extract RNA using one of the various methods. We used phenol-chloroform extraction. Tested quality by NanoDrop spectrophotometry and on some of the samples also the Agilent BioAnalyzer (automated electrophoresis).

**Isolation of mRNA**

* **Why**: Total RNA extracted from cells contains various types of RNA, including ribosomal RNA (rRNA), transfer RNA (tRNA), and mRNA. For mRNA sequencing, we're primarily interested in the mRNA fraction.
* A screenshot of a computer

  Description automatically generated**How**: mRNA molecules have a polyadenylated tail (poly-A tail) at their 3' end. By using magnetic beads coated with poly-T oligonucleotides, mRNA can be selectively bound and isolated from the total RNA, as these oligos can hybridize with the poly-A tail. Because an intact poly-A tail is needed, it is not ideal for degraded samples.

Total RNA contains a lot of ribosomal RNA (rRNA), which isn't usually of interest in mRNA sequencing. To get just the mRNA, you can use poly-T oligo-attached magnetic beads. These beads bind to the poly-A tail found at the end of mRNA molecules. rRNA depletion is another method.

A diagram of a sequence of dna

Description automatically generatedWe used the Illumina stranded mRNA prep kit. This is designed for the preparation of sequencing libraries[[1]](#footnote-1) from mRNA. The “stranded” aspect means that the kit preserves the information about which DNA strand the RNA was transcribed from. This is valuable because it allows for the determination of the directionality of transcription which can be crucial for understanding gene regulation, alternative splicing, and other transcriptomic phenomena.

A diagram of a diagram

Description automatically generated with medium confidencePoly-T Oligo attached magnetic beads works sine mRNA molecules have a unique features: a polyadenylated tail (poly-A tail) at their 3´end. This tail can be targeted for mRNA isolation. Beads coated with poly-T oligonucleotides are used. Hese can hybridize bind to the poly-A tail. When total RNA is incubated with these beads, the mRNA binds to the beads due to the complementary base pairing between the poly-A tail and the poly-T oligos. After binding, the beads (with the attached mRNA) can be separated from the rest of the RNA using a magnet. The mRNA can then be eluted from the beads, giving a purified mRNA sample.

**Library preparation and reverse transcription**

With the Illumina stranded mRNA prep kit, fragmentation is done before reverse transcription.

* **Fragmentation & Priming:** Once mRNA is isolated, it is fragmented into smaller pieces. The fragmented mRNA is then primed for the synthesis of the first-strand cDNA.
  + **Why:** mRNA molecules can be quite long, and sequencing machines read short fragments.
  + **How**: This is usually done using heat and divalent cations or enzymatic methods.
* **First-Strand cDNA Synthesis**: Using reverse transcriptase, the fragmented mRNA is converted into its complementary DNA (cDNA) strand.
  + **Why:** Sequencing machines cannot read RNA directly. Instead, they read DNA. Thus, the mRNA fragments need to be reverse transcribed into complementary DNA (cDNA).
  + **How:** Using an enzyme called reverse transcriptase and random primers, the mRNA fragments are converted into their cDNA counterparts.
* **Second-Strand cDNA Synthesis**: This step incorporates dUTP in place of dTTP, which is crucial for the strandedness of the library. The second strand synthesized will be complementary to the first strand and will be used as a template in subsequent PCR amplification.
  + **Why:** To create double-stranded DNA, which is more stable and is required for the subsequent steps.
  + **How:** This step involves synthesizing the second strand of the cDNA. In stranded mRNA sequencing, dUTP is incorporated during this step to ensure strand specificity later in the process.
* **End repair and A-tailing:**
  + **Why**: The fragmented cDNA might have jagged ends. For efficient adapter ligation, these ends need to be smooth. Additionally, a single 'A' nucleotide is added to the 3' ends to prevent them from ligating to one another and to prepare them for adapter ligation.
  + **How**: Enzymatic treatment is used to polish the ends and add the 'A' nucleotide.
* **Adapter Ligation**:
  + **Why:** Adapters are ligated to the cDNA fragments. These adapters are necessary for the fragments to bind to the flow cell of the sequencing machine, provide a starting point for PCR amplification, and contain indices for sample identification (since we are multiplexing samples).
  + **How**: Adapters are ligated to the prepared cDNA fragments using a ligase enzyme.
* **PCR Amplification**: The library is then amplified using PCR to generate sufficient quantities for sequencing. Before sequencing, a PCR amplification step is performed. The enzyme used in this PCR step does not recognize dUTP, so the second cDNA strand (which contains dUTP) is not amplified. This means that only the first-strand cDNA, which is complementary to the original mRNA, is amplified and subsequently sequenced.
  + **Why:** To ensure there's enough material for sequencing, the library is amplified.
  + **How:** PCR is used to selectively amplify the adapter-ligated fragments. This step also incorporates unique indices for sample multiplexing.
* **Library validation:**
  + **Why:** Before sequencing, its essential to ensure the library is of good quality and the right size.
  + **How:** Techniques like qPCR, electrophoresis, or using instruments like the Agilent Bioanalyzer or TapeStation can be used.
* **Strand-Specific Sequencing**: Due to the incorporation of dUTP during the second-strand synthesis, only the original mRNA strand (first-strand cDNA) is sequenced. This ensures that the directionality (strand information) of the original mRNA is preserved.
* **Pooling & Sequencing:** If multiple samples are being sequenced together, the individual libraries (each with a unique index) are pooled in equal proportions. The pooled library is then loaded onto the sequencing machine.

**Sequencing**

The Illumina NovaSeq 6000 utilizes a sequencing technology called "sequencing by synthesis" (SBS).

**Data analysis workflow**

**Generation of FASTQ files**

* Use software bcl2fastq
* Requires a linux/UNIX environment.
* A command line tool to **convert** **proprietary cbcl files to the more standard FASTQ format**, along with performing tasks like **demultiplexing** samples based on index sequences.
* Use a virtual UNIX environment, or ucloud which is made for heavy computational tasks.

The NovaSeq 6000 outputs data in as highly compressed .cbcl (compressed base call) files. This format is designed to store the nucleotide base calls from sequencing runs and their corresponding quality scores in a compressed manner. This is a binary format which contains only the base calls and the quality scores, excluding other metadata. Given that cbcl is a proprietary format, it's generally necessary to convert these files to the more universally usable FASTQ format for downstream bioinformatic analyses. This is usually done using Illumina's bcl2fastq conversion software.

A screenshot of a computer error

Description automatically generatedA screenshot of a computer

Description automatically generated

**FASTQ files quality control**

* **Objective**: The goal is to evaluate various quality metrics, including per-base quality scores, GC content, and adapter content, among others.
* **Software**: Tools like FastQC and MultiQC are commonly used for this purpose.
* **Output**: You'll get a series of plots and metrics that give you an overview of your data quality.
* **Decision-making**: Based on these QC reports, you decide whether the data is of sufficient quality for downstream analysis or if it requires pre-processing steps like trimming or filtering.

Typical QC Metrics

* **Per-base sequence quality**: Provides quality scores for each base across all reads.
  + Measure of the quality of the nucleotide calls for eeach base position in the sequencing reads. Ie., how certain the machine is that it called the right base. If all reads in the cluster calls the same base it is very certain. If different bases are called it is uncertain.
  + Low quality bases may introduce errors in downstream analyses like alignment and quantification. It’s common to trim or filter reads based on quality scores.
  + High Quality: Scores remain high throughout the read length, generally indicating good data.
  + Decreasing Quality: A gradual decline, particularly toward the end of reads, is common and often results from the limitations of sequencing technology.
  + Remedial Actions:
    - Trimming: Bases with low quality at the read ends can be trimmed off.
    - Filtering: Entire reads with average quality below a certain threshold can be removed.
* **Per-sequence GC content**: Checks if the GC content distribution conforms to expected values.
  + The GC content of each read is calculated, and the frequency distribution of these GC contents is plotted against the expected distribution.
  + Abnormal GC content can indicate various issues such as sequencing bias or contamination.
  + A histogram depicting the GC content distribution. A bell-shaped curve is often seen as indicative of good quality, while a skewed or multi-modal distribution can be problematic.
  + Normal Distribution: A bell-shaped curve indicates that the sample likely represents the transcriptome accurately.
  + Skewed/Bimodal Distribution: This could suggest contamination or an unnatural enrichment of certain transcripts.
  + Investigate Source: If GC content is off, check sample preparation steps, and you may need to sequence a new sample.
  + Normalization: Some advanced normalization methods can partially correct for GC bias, though this is generally not recommended as a first line of action.
* **Overrepresented sequences**: Identifies any sequences that appear more frequently than expected, which could indicate contamination or adapters
  + All reads are compared against each other to identify duplicates or highly similar sequences. The frequency of each unique sequence is calculated.
  + Overrepresented sequences could indicate multile issues like adapter contamination, OCR duolicates, or even specific kinds of sequencing errors. These need to be identified and potentially removed before downstream analyses.
  + No Overrepresentation: Ideal case, suggesting no contamination.
  + Presence of Overrepresented Sequences: Could indicate adapter sequences, PCR duplicates, or contamination.
  + Remedial Actions:
    - Adapter Trimming: Removalof adapter sequences.
    - Duplicate Removal: Software like Picard can remove PCR duplicates, although this is more common in DNA-seq than RNA-seq.

A screenshot of a screen

Description automatically generatedA graph with a red line

Description automatically generated

A graph showing a number of lines

Description automatically generated with medium confidenceA graph of a normal distribution

Description automatically generatedA screen shot of a graph

Description automatically generated

**Preproccessing FASTQ files**

Preprocessing is the next crucial step after initial Quality Control (QC) checks. The primary aim here is to clean and prepare the raw sequence data for accurate downstream analysis. BBDuk (Bushnell B. Decontamination Using Kmers) is used by the bioinformatician at the genetics department.

Workflow:

1. **Objective**: The main goal is to remove undesirable elements from your sequence data, such as low-quality bases, adapters, and other contaminants that can skew your results.
2. **Common Functions**: Adapter trimming, quality trimming, k-mer-based filtering, and length filtering.
3. **Input**: Your raw FASTQ files or the files post-initial QC, depending on your workflow.
4. **Output**: Cleaned FASTQ files that are ready for read alignment and other downstream analyses.

Typical BBDuk Operations:

* **Quality Trimming**: Removes low-quality base calls from the ends of reads. Removal of reads with a low mean phred-score.
* **Adapter Trimming**: Detects and removes any adapter sequences.
* **K-mer Filtering:** Allows for the removal of sequences based on specific k-mers. This can be useful for removing contaminants.
* **Length Filtering**: Removes sequences that are below a certain length.

Explanation of K-mer filtering:

**Objective:** The primary aim is to filter out sequences based on their k-mer composition, where a k-mer is a contiguous subsequence of “k” nucleotides from a read.

**Mechanism:** K-mer filterering typically involves scanning each read for the presence of specific k-mert that we want to remove or keep.

**Use cases:** This is valuable for removing contaminats or sequences that should not be part of our analysis, like secific bacterial sequences in a human RNA-seq dataset.

**Output:** The output is a sequence of reads devoid of the undesirable k-mers, making the dataset cleaner and more focused on the target of interest.

In BBDuk we create a library of k-mers that we want to filter out. This can come from known contaminants, adapters, or any sequence we believe should not be in our dataset. Each read in our FASTQ files are then scanned for the presence of these k-mers in our library. Read containing the flagged k-mers are either removed from the dataset or truncated to eliminate the k-mer. The presence of contaminants or irrelevant sequences can introduce noise and confounders into my analyses. Using k-mer filtering helps ensure that the data is as precisce and relevant as possible. An example: if working with human hepatocytes but finding bacterial contamination, k-mer filtering can help remove reads specifically attributable to the bacterial genome, thereby improving the quality of your human-specific analyses.

We are working with human hepatocytes, but there's a chance that bacterial contamination exists in our samples. We want to filter out these bacterial sequences.

* 1. Create k-mer library
     1. The bacterial species has a unique k-mer sequence: "AGCTGGA," and we don't expect to find this 7-mer in human sequences.
  2. Our FASTQ files has millions of reads, and some contain the bacterial 7-mer "AGCTGGA."

Read1: ATCGGGGAGCTGGATCGA

Read2: AATTCCGGAGCTGGAGCT

Read3: TTAGCTAGCTGGATAGC

Read4: ATCGGATCGAACTGACG

* 1. BBDuk scans each read in the FASTQ files, looking for the presence of "AGCTGGA."

Read1: ATCGGGGAGCTGGATCGA

Read2: AATTCCGGAGCTGGAGCT

Read3: TTAGCTAGCTGGATAGC

Read4: ATCGGATCGAACTGACG

* 1. Reads containing the 7-mer sequence would be flagged. Here read 1, 2, and 3.
  2. We can either remove flagged reads, or remove just the part of the read containing the flagged k-mer.

**Aligning our FASTQ files with a reference transcriptome**

Objectives and basics

1. **Objective**: To map the processed sequencing reads onto a reference transcriptome, which helps in quantifying expression levels and identifying splice variants.
2. **Alignment Tools**: There are various tools available for this, such as STAR (used by us), Hisat2, and Bowtie2. The choice of tool often depends on your specific requirements and computational resources.
3. **Input**: Your cleaned and preprocessed FASTQ files.
4. **Output**: Typically, the alignment output is in SAM/BAM format, containing detailed information on how each read aligns with the reference.

Steps for Read Alignment

1. **Indexing**: Before alignment, the reference genome or transcriptome usually needs to be indexed by the chosen alignment tool. This makes the actual alignment process more efficient.
2. **Alignment Execution**: Run the alignment program, specifying the reference index and the FASTQ files as inputs.
3. **Quality Check**: After alignment, it's good practice to evaluate the quality of the alignment using various metrics, such as the percentage of reads that were successfully aligned.

Reference genomes/transcriptomes for alignment can be found at the database: <https://www.ensembl.org/index.html>.

We used the release 108 of the human transcriptome (Homo\_sapiens.GRCh38.108.chr.gtf.gz).

**Quantification**

After alignment the next step is quantification. We have used HTseq for quantification.

RNA-seq quantification aims to translate aligned reads to numerical values indicative of transcript abundance. This write-up delves into the role of HTSeq in obtaining raw counts and the concept of Counts Per Million (CPM) for normalization.

HTSeq for Quantification

1. **Role of HTSeq**: HTSeq is a Python-based tool that is specialized in producing a raw count matrix based on aligned reads. It generally works with SAM/BAM files as input. The aligned reads are used to count the number of reads mapped to each gene or transcript, which serves as a raw measure of its expression level.
2. **Workflow**: HTSeq produces raw counts that serve as the basis for downstream analyses such as normalization and differential gene expression analysis.

Transition to Normalization

1. **HTSeq to R**: Once raw counts are generated using HTSeq, this count data usually transitions to an R environment for normalization and other downstream analyses.
2. **R Packages**: Various R packages, such as DESeq2 and edgeR, offer methods for normalization and differential expression analysis.

CPM: Counts Per Million for Normalization

1. **What is CPM**: Counts Per Million (CPM) is a straightforward method of normalization. It scales the raw read counts by the total number of reads, making it easier to compare expression levels across samples.
2. **Formula and Interpretation**:
3. **Rationale**: Normalization methods like CPM correct for varying sequencing depths and make gene expression comparable across samples.

**Impact of Sequencing Depth on Non-Normalized Expression**

1. **Magnitude of Counts**: Sequencing depth refers to the number of reads that are generated for a sample. Higher sequencing depth generally results in higher raw read counts for all genes, not just those that are differentially expressed. If one sample has a higher sequencing depth, genes may appear to have higher expression in that sample simply due to this imbalance.
2. **Statistical Power**: A higher sequencing depth can increase the statistical power of your analysis, allowing you to detect smaller changes in gene expression. However, this doesn't automatically mean that a gene is more biologically relevant just because its expression level appears higher.
3. **Sensitivity and Specificity**: Greater sequencing depth can also identify rare transcripts but could lead to false-positive identifications if not properly normalized.
4. **Normalization Necessity**: Without normalization, comparing raw counts between samples can lead to misleading results. It might seem like genes are differentially expressed when, in reality, the differences arise from variations in sequencing depth.

Sequencing depth is fundamentally tied to the number of reads generated, which in turn is influenced by the amount of RNA present in each sample. While sequencing depth can provide insights into the relative abundance of different transcripts, it can also introduce bias if not properly accounted for.

Why Sequencing Depth Can Be Misleading:

Variation in RNA Input: The amount of total RNA in each sample can vary due to differences in cell numbers, RNA extraction efficiency, or other experimental factors. If one sample has more total RNA, it will naturally generate more reads, making genes appear more highly expressed.

Denominator in Normalization: The total count of reads in a sample serves as the denominator when calculating metrics like CPM (Counts Per Million). If one sample has a higher overall read count, all genes in that sample will have proportionally higher raw counts, potentially skewing the analysis.

Biological vs. Technical Variation: The challenge lies in distinguishing between biological changes in gene expression (e.g., due to inflammation in hepatocytes) and technical variations (e.g., more RNA in one sample). Normalization methods aim to eliminate or minimize these technical biases.

Differential gene expression analysis (DGEA) enables the identification of genes that exhibit variable expression levels between two or more conditions, e.g., cases vs. controls or treated vs. untreated.

Biological Interpretation: Tools like GSEA (Gene Set Enrichment Analysis) or DAVID can be used for functional enrichment analyses to provide biological context to your list of differentially expressed genes.

<https://www.cs.jhu.edu/~langmea/resources/lecture_notes/dna_sequencing.pdf>

<https://data-science-sequencing.github.io/>

<https://www.youtube.com/watch?v=tlf6wYJrwKY>

1. A library is simply a collection of DNA fragments that have been prepared for sequencing. [↑](#footnote-ref-1)