Module 2.1: Learn - Biology

2.1: Learn Biology/Statistics

Experimental design for a differential gene expression experiment

In this section, we wanted to take some time to explain to you how an omics-level differential expression experiment works. This differential expression experiment aims to measure the expression of all genes across two or more groups of samples with the intention of figuring out (a) which genes are expressed at particularly higher or lower levels (differentially expressed) in a specific group of samples compared to another or (b) follow a specific trend across samples. The reason we do this is that we want to figure out which genes are involved with specific phenotypes or conditions. If you figure out what genes are changing in a particular condition, you can often look up the function of those genes and come up with a hypothesis for how those genes are involved in the phenomenon you are studying and then design follow-up experiments to test your hypothesis.

RNA sequencing for relative expression differences

Having learned about RNA sequencing in Module 1, we recall that gene expression is quantified using the total number of reads aligning to a gene’s coordinates. We load the same amount of starting material for each sample when creating RNA sequencing libraries, so that way we know that differences in the number of reads for each gene are not due to differences in the amount of input material. In this way, RNA sequencing gives us a *relative* estimation of gene expression across many samples in a study. The number of counts you get does not correspond to any absolute measure of gene expression.

Sample Number and Replicates

In order to calculate statistics that allow us to identify gene expression changes that are unlikely to happen by chance, we need to have at least three independent samples per group. For the samples used in the experiment for this course, each placenta collected represents one independently acquired sample and we have 10 females and 12 males. Independently acquired samples in the same group are called *biological replicate* replicates of female placentas, 12 biological replicates of male placentas, total number of samples is 22).

For each placenta, two tissue samples from opposing quadrants were taken for RNA sequencing. Samples taken placenta represent *technical replicates* as they are not independent from one another. Having technical replicate robustness of your results because taking sums or averages of multiple measurements helps to reduce the effec that you are not trying to identify in your study. In our differential expression pipeline, we will take the sum of the t per placenta

Prioritizing candidate genes

With the advent of whole-genome (or whole-transcriptome) level assays like RNA sequencing, we are able to me of virtually all genes at once, giving us an unsupervised, data-driven method to discover genes' whose changes i how a specific cellular behavior takes place. As great as it is to let our data tell us which genes are important ins biases, the challenge then is to prioritize genes according to the likelihood that they are responsible for true biolo between sample groups, in our cases between female and male placenta tissues. There are several key aspects expression that we use to prioritize candidate genes.

When there are enough biological replicates to conduct a statistical test of differential gene expression, this is a k allows us to identify genes with a large fold change difference between sample groups being compared and low v samples within the same group. There are many different types of statistics we can use that make different assu you are comparing (see additional resources for more information about different statistical tests); we can choose for our studies. Because we are doing many measurements as RNAseq measures expression of every gene, we our statistics to account for the increased chance of finding statistically significant results with a high number of m come on this in future modules.

Another way to prioritize candidate genes is by looking at what has been discovered about that gene in previous case, we might take special interest in genes that have been studied other placenta studies, have shown sex diff experiments, have been associated with functions of the placenta such as nutrient transport or gas exchange, or play a role in pregnancy complications.

There are many ways you can look up information about genes identified to be differentially expressed, here are

1. Published papers that talk about the gene
    - **Google Scholar (https://scholar.google.com/)** will use Google search engine to mine published papers, reference materials
    - **Pubmed (https://pubmed.ncbi.nlm.nih.gov/)** is the official databank of published papers
        - Searching for a gene will give you a list of abstracts for papers that contain the search terms you enter for full text of the paper
        - Make sure you are logged into the ASU network because ASU has paid for access to many journals sc whole paper instead of just the abstract
    - Gene information databases
        - **Genecards (https://www.genecards.org/)** is an awesome tool that tells you all kinds of information a their known function, genomic coordinates, disease the gene has been associated with, and other nam gene (can help you search for more information)
    - Tissues where the protein product of the gene is expressed
        - **The Protein Atlas (https://www.proteinatlas.org/)** is a resource that allows you to search for the prot in a variety of cell types, cell lines, and human tissues in health and disease
            - We might prioritize genes whose protein product has been detected in the placenta or other pregna
    - Interaction networks
        - When you identify multiple genes that are differentially expressed, we might be interested in figuring ou interact with each other in some way
            - **STRING (https://string-db.org/)** is a resources that tries to find connections between genes
                - Shows interactions of many different types
                - Displays information about the function of the gene(s) you enter and their predicted interaction

Effects of trimming on biological inference of differential expression

Our research project asks how trimming of sequencing reads affects differential gene expression analysis. To un more clearly, we need to understand what trimming software (bbduk in our case) does. In Module 1, we learned is performed: mRNA transcripts from a tissue sample are transcribed to cDNA, fragmented, copied, and sequenc RNA sequencing library preparation kits use specific adapters to sequence the short reads which include fragme Because polymerases tend to get less accurate as the read length increases, the end of sequencing reads tend t

quality reads, as defined by the PHRED score which are metrics given to assess the quality of the nucleotide ide
sequencer.

Trimming software does two types of trimming. The first is adapter trimming; sequencing adapters are supposed
of the RNA sequencing library preparation protocol, but this process is not perfect so some of the sequences rem
Adapter trimming is done by bbduk by removing sequence aligning to Illumina adapter sequence (which we provi
second type of trimming is quality trimming. For this, trimming software slides a window down the length of the re
removes areas that have higher amounts of nucleotides given a PHRED score below the threshold that we provi
in bbduk). We will vary this threshold along with the threshold for how much of the read is left after trimming (`mi`
bbduk).

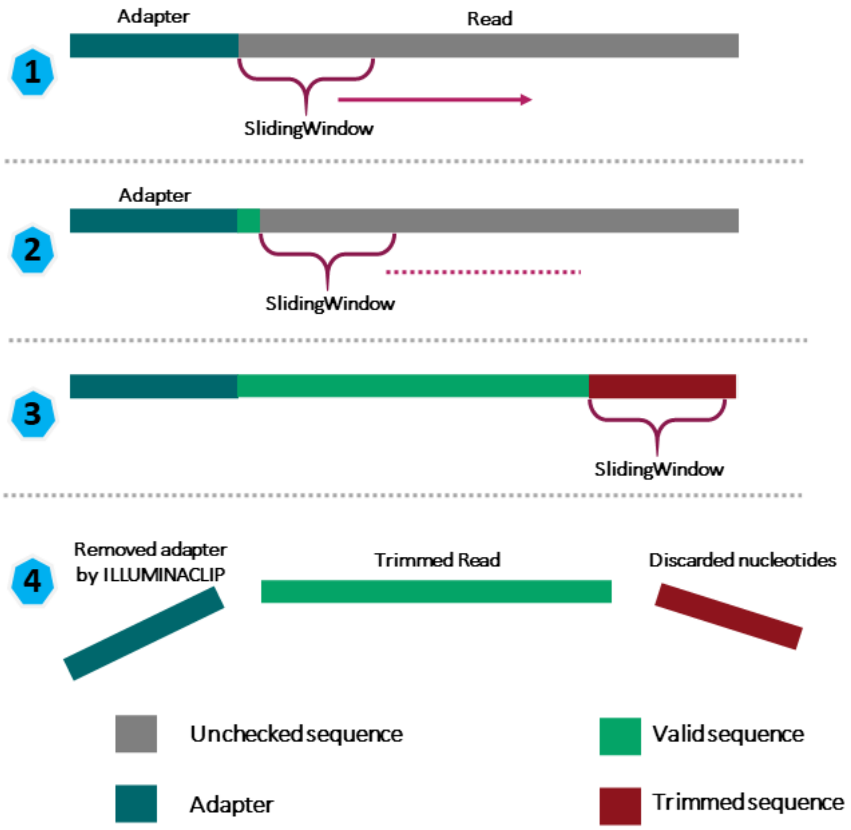

Figure: **Trimming of RNA sequencing reads.**
mRNA transcripts from tissue samples are transcribed to cDNA, fragmented, copied, and sequenced in short reads. (1) RNA sequencing library preparation kits use specific adapters to sequence the short reads which include fragments from the cDNA. (2) The sliding window cuts if the average quality within a group of bases falls below the specified threshold. (3) Because polymerases tend to get less accurate as the read length increases, the end of sequencing reads tend to have more low quality reads and are discarded. (4) The resulting trimmed

read is a valid sequence with the adapter and low quality ends removed. (**Source** (https://carpentries-incubator.github.io/metagenomics/03-trimming-filtering/index.htı

Depending on what values of trimming parameters are supplied, we will get different sets of trimmed reads remai below shows how those reads are mapped to genes during the alignment step. Alignment can be done to the ge (which is what we did to generate the data for this course) or to transcriptome sequence if you are looking for rea sites.

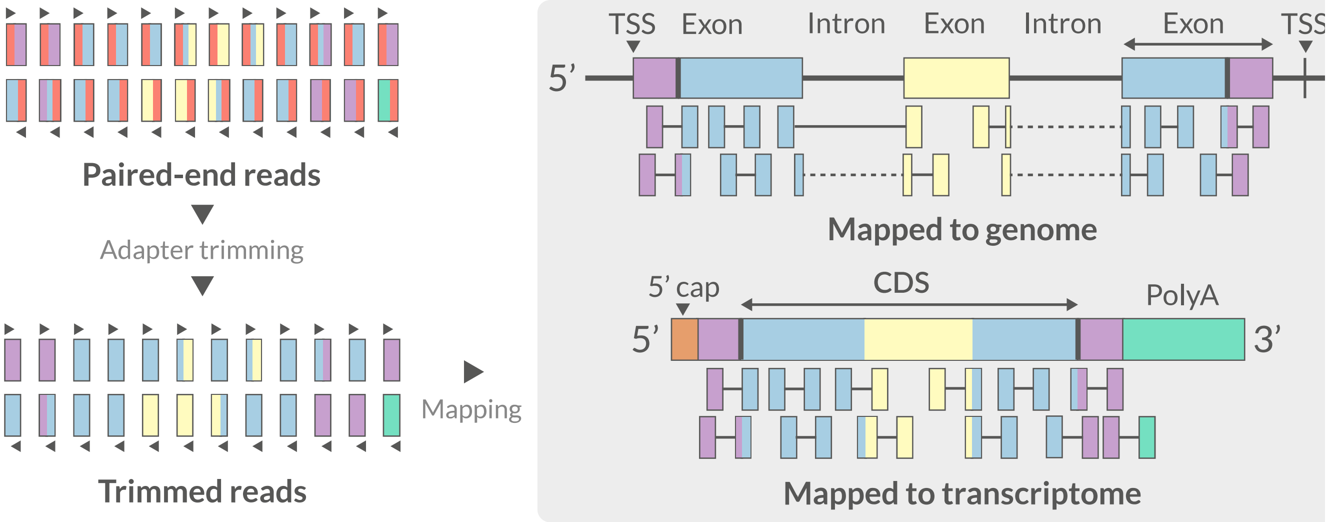

Figure: Schematic of alignment of RNA sequencing reads after adapter trimming. Paired-end reads have the adapters and specified parameters trimmed then are mapped to the reference genome sequence or the transcriptome sequence during alignment. (**Source** (https://nbisweden.github.io/workshop-ngsintro/2001/slide_rnaseq.html#18))

Once alignment is complete, we use software called featureCounts to assign reads to genes. Given the alignme reads to the reference genome (.bam file) and an annotation file continaining the coordinates of all genes (includi introns are), featureCounts quantifies a gene's expression as the total number of reads uniquely mapping to it. T how featureCounts counts reads that align to any feature of the gene and allows for insertions, deletions, exon-e fusions.

Different trimming parameters will lead to shorter or longer read sequences and may result in some reads being f contain too few base pairs after trimming. Since genes are quantified based on the number of reads assigned to affect the number of reads we get that will be mapped to genes and counted with featureCounts. If the quantificat sample is changed, we could see differences in the average expression of genes in a sample group and thus affe expression between sample groups being compared (females versus males in our case).

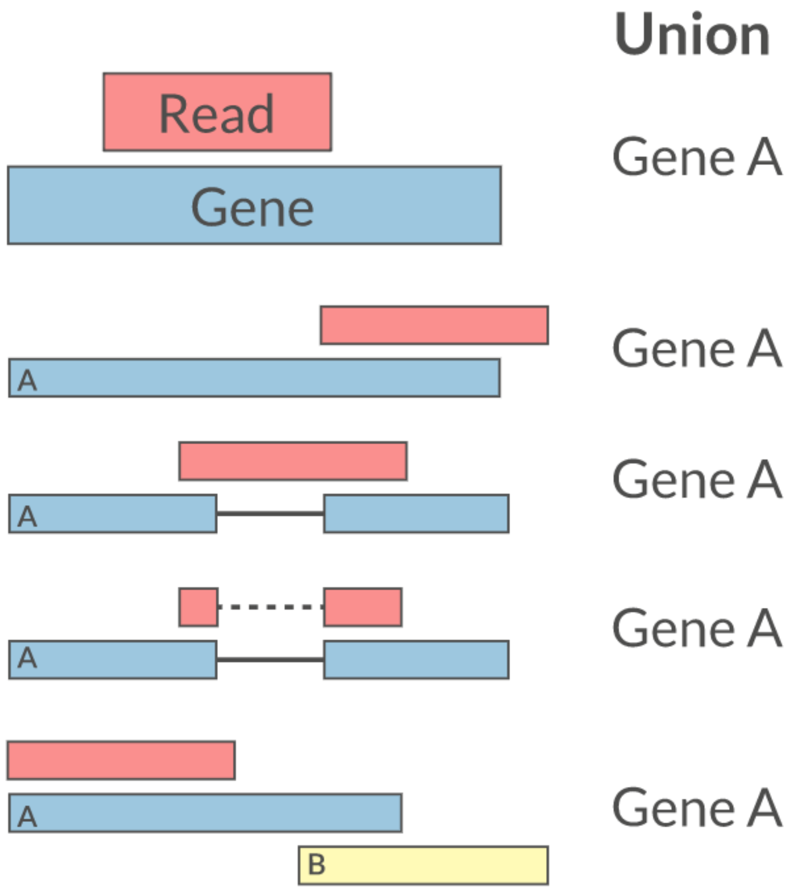

Figure: **How reads are assigned to genes when taking counts for quantification.** A sequencing read is represented in red and the sequence of a gene in the reference genome is in blue. Read counts are indicative of gene expression and reads can be quantified on features such as

genes or transcripts.
(**Source** (https://nbisweden.github.io/workshop-ngsintro/2001/slide_rnaseq.html#32))

Module 2.1 Additional Resources

- **Differential Expression Experimental Design** (https://ucdavis-bioinformatics-training.github.io/2018-June Workshop/tuesday/ExperimentalDesign.pdf)
- **Workshop on analysis of bulk RNAseq** (https://nbisweden.github.io/workshop-ngsintro/2001/slide_rnaseq
- **Choosing the Right Statistical Test | Types & Examples** (https://www.scribbr.com/statistics/statistical-tes
- **Lesson on trimming and filtering of sequencing data** (https://carpentries-incubator.github.io/metagenom filtering/index.html)