Module 1.1: Learn - Biology

Overview

By the end of this module, you should be able to explain the general biology of the placenta, why we are interested in placental genes that have higher expression in one sex versus the other, the difficulties involved in studying genes on the sex chromosomes which are likely differentially expressed between the sexes, and have a thorough knowledge the data we will be working with. By the end of this course, you will have conducted differential gene expression analysis of real human placenta samples.

If you have questions about any of the content, please post in the course slack under the channel for module 1.

In this section we will work on:

1. Course description and research question
2. Placenta biology
3. Sex differences and sex chromosomes
4. Genomics/transcriptomics refresher
5. Gene expression and RNAseq preprocessing

Research Aims

What is this course?

This is a course-based research experience specifically designed to give ASUOnline students the opportunity to attain research experience, gain technical skills, and give back to the scientific community. The Genomics Research Experience is a 7-week research experience that answers *real* research questions with *real* research techniques. Notice the emphasis on the words real and research. This means we do not know the outcome of our research question and will be working as a team to identify best practices, evaluate, and

troubleshoot. The results we produce will be used in a real publication, so please be thorough in your understand taking. Further, because this is an authentic research experience – meaning we have not conducted these analys things could go wrong. In fact, as in all authentic research experiences, we will get both expected and unexpecte analyze those results together.

Each module contains three sections that are aimed to help you learn (1) the **biology** behind our research, (2) th to perform our research, and (3) the **professional development** skills used in real genomics research. Let's take question we are seeking to answer throughout the duration of this course.

The research question

In this research experience, we are asking a very specific question:

> **What are the effects of trimming next-generation sequencing reads on the measurement of sex differences in gene expression in the human placenta?**

For any computational analysis, researchers usually choose data processing parameters based on what they fin groups in publications or what was used successfully in their own past studies. What was chosen for those data p may (or may not) have big effects on findings and conclusions drawn from that analysis. In this project we are ex parameters for trimming, a very common early data processing step for next-generation sequencing analysis. Fi will be widely relevant since many groups use trimming as the first step in their sequencing data analysis workflo about the basic steps of data processing and differential gene expression analysis in this course.

The Wilson lab at ASU is interested in studying the biology of the placenta, a crucial organ in human developmer human placenta samples and determined a set of genes that had significantly higher expression in one of the se

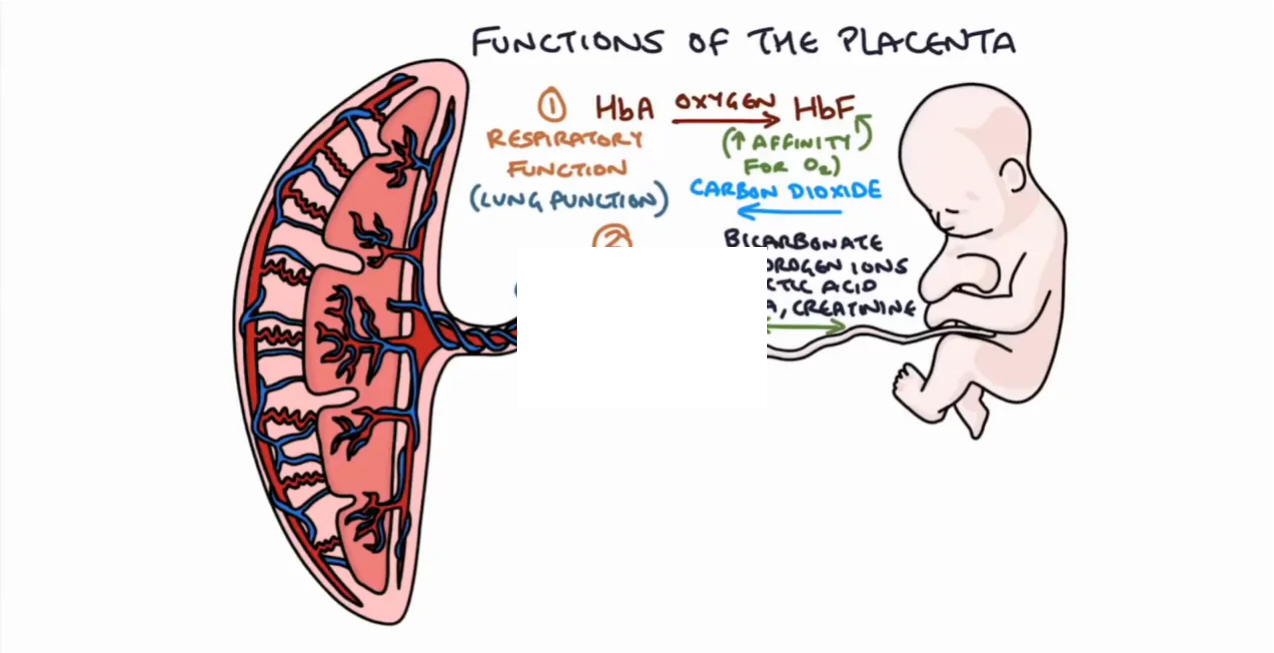
males) using specific data processing parameters, but we wanted to know how robust our results are. Since the differentially expressed between the sexes in the human placenta will serve as the test for effects of varying trimr Having worked with this data and these samples in the lab, we have already established that this data set is biolc high quality, and does in fact show a signal for gene expression differences; this establishes a baseline for this re

Placenta biology and sex differences

Placenta Biology

To give you some background on why we chose to use genes differentially expressed by sex in the placenta as a effects of trimming, we will start by describing the function of the placenta and talk about what is known about ge tissue.

The placenta is a highly specialized organ that develops in the uterus during pregnancy serving as an interface b and the growing fetus. The **placenta (https://www.mayoclinic.org/healthy-lifestyle/pregnancy-week-by-week/in-d 20044425#:~:text=The%20placenta%20is%20an%20organ,umbilical%20cord%20arises%20from%20it.)** nourishes, pr removes waste from a developing fetus. The video below gives an overview of the structures of the placenta and



Video. Understanding the Placenta.

This video succinctly reviews the placental structures and the five key functions of the placenta.

View Transcript. (https://canvas.asu.edu/courses/122165/files/54792578?wrap=1) **(https://canvas.asu.edu/courses/122165/files/54792578/download?download_frd=1)**

The placenta has two sides – one is connected to the fetus by the umbilical cord and the other is connected to th maternal tissue called “decidua”. Contained within the umbilical cord are two arteries that carry deoxygenated blc fetus and a single umbilical vein that carries oxygenated blood to the fetus.

Within the placenta itself, there is a network of umbilical and maternal veins and arteries. The umbilical arteries a the placenta form tree-like structures surrounded by the placental membrane; in between is the intervillous space exchange between maternal and fetal blood. Maternal blood forms pools outside the placental membrane and m

between the mother and fetus and across the membrane without mixing the blood. This process of diffusion allow
of the placenta.

Figure. The structures of the placenta.

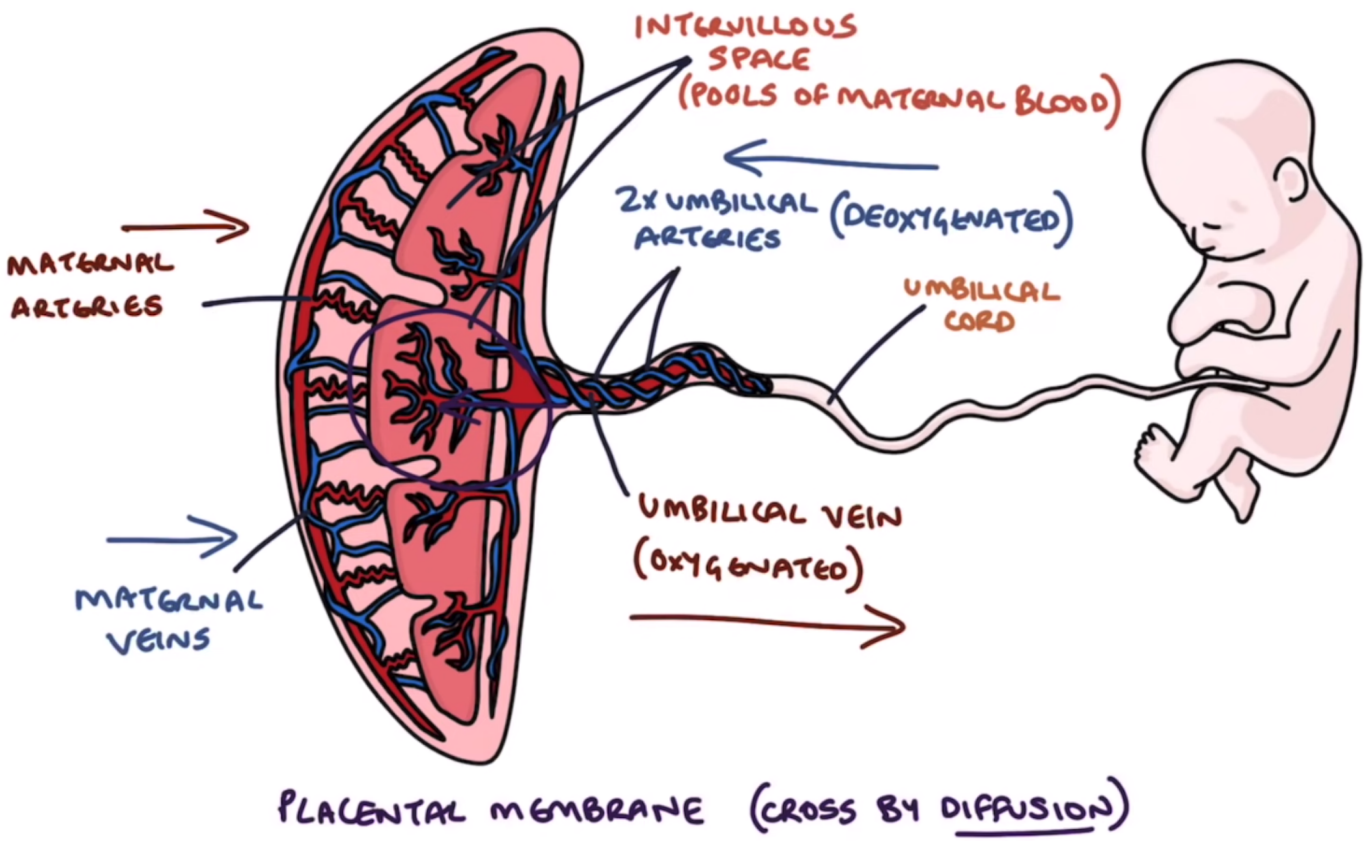


Figure. The structures of the placenta.

The main structures of the placenta include the arteries and veins on the maternal side, the intervillous space as the site of exchange, the two umbilical arteries for deoxygenated fetal blood, and the umbilical vein for oxygenated blood to get sent to the fetus.

The five key functions of the placenta

1. **Respiratory function:** allows the maternal blood to exchange oxygen with the fetal blood and for CO_2 waste the fetal blood
2. **Excretion function**: rid fetus of waste and provide osmoregulation
3. **Nutrition function**: carbohydrates, vitamins, and nutrients from maternal blood diffuse across the placental m blood stream
4. **Immunity function**: maternal antibodies from previous infections can cross the placental membrane to protec pregnancy and temporarily after birth
5. **Endocrine function**: the placental tissue creates hormones that help maintain the pregnancy, such as HCG (gonadotropin), estrogen, and progesterone

Sex differences in the placenta and pregnancy

In this research, we are comparing gene expression differences between placentas from individuals with two X cl placentas from individuals with an X and a Y chromosome (XY). Recall that the XX individual receives one X chro maternal genome and the other from their paternal genome, while the XY individual receives the X from the mate Y from the paternal genome. As the placenta and fetus are derived from the fertilized egg, they share the same g XX fetus will develop an XX placenta and an XY fetus develops an XY placenta.

In the past, the placenta had been viewed as an asexual organ and that resulted in biased studies that didn't con placental function. This is especially important in complicated pregnancies where the maternal condition combine differences of the placenta impact the development of the fetus. For example, sex differences have been observe characteristics like birth weight (XY males tend to be larger) or pregnancy complications (XY males more often ex outcomes) and these outcomes can be intensified when the mother has a medical complication such as hyperter

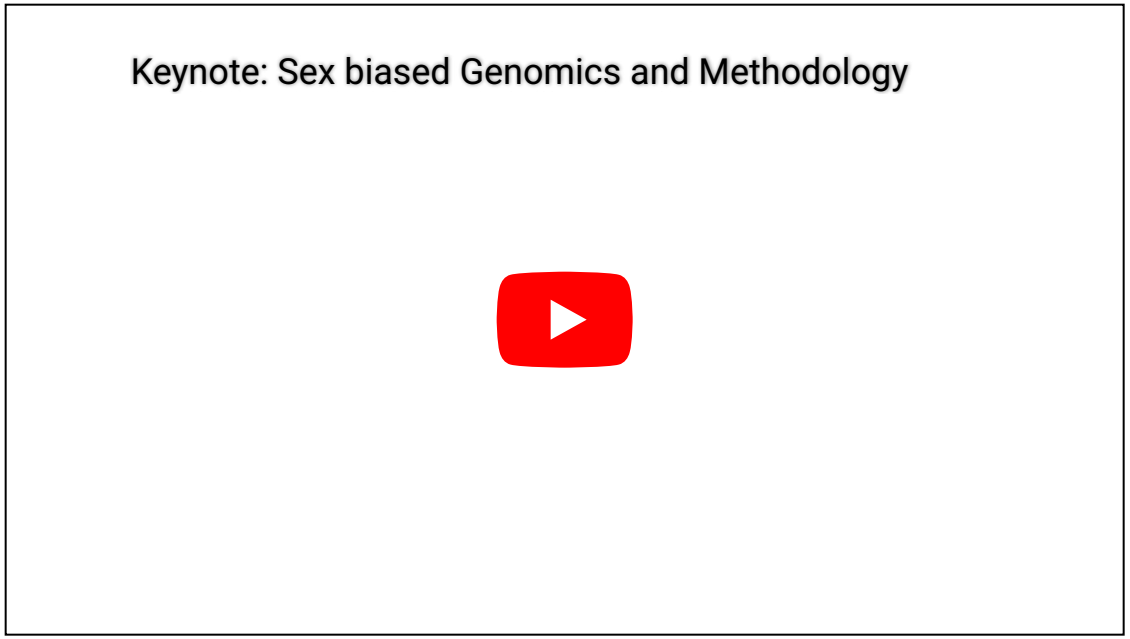
One study on pregnancies complicated by asthma found that if the mother went untreated, the XX fetus was sma had the asthma under control, the XX fetus appeared to be unaffected; conversely, the XY fetus was unaffected b either condition. Another study observed that sex differences may even play a role in the first 48 hours of a prema survival rate as they are born more vasodilated than the premature XX individuals. Furthermore, sex differences known to impact the immunity function of the human placenta and sex-specific fetal responses to complicated ma that contribute to these adverse outcomes (Clifton, 2009).

In our data set, we aimed to have our sample set have as close to a 50:50 ratio of XX:XY as possible. We were a
analyze ten (10) XX and twelve (12) XY samples (22 total). Our research focuses on sex differences in placental
health so that we can better understand how changes can lead to sex differences in poor outcomes for the fetus,

Sex chromosomes

To help you learn about and communicate results about sex differences, it is important to use the appropriate lan
discussing them. We are interested in genetic differences due to genetic sex (sex chromosome genotype) – not c
with gender, which are based on how a person self-identifies. The sex chromosome complement is the set of sex
contained in the nucleus. Although genetic males typically have one X chromosome and one Y chromosome, an
typically have two X chromosomes, the number of X and Y chromosomes can vary. Common sex chromosome v
Kleinfelter syndrome (47,XXY) and Turner syndrome (45,X). Further, sex chromosome content can vary over tim
individuals can lose their Y chromosome within a proportion of their cells over time as they age; loss of Y chromo
observed in cancer cells.

As part of the data processing for this research, we take one extra step - compared to typical genomics pipelines
to assess gene expression from sex chromosomes. This talk was featured in a meeting from Bioconductor which
source software project that uses the R statistical programming language; you will be using Bioconductor packag
remainder of this course. The video is included below as supplementary viewing.



Video. Bioconductor Conference 2021, Keynote Speaker: Melissa Wilson, PhD.
In this video, Dr. Melissa Wilson, Computational Evolutionary Biologist at ASU, presents *Sex-biased genome evolution* to discuss the structure, function, and evolution of the sex chromosomes.
View Transcript. (https://canvas.asu.edu/courses/122165/files/54912778?wrap=1) **(https://canvas.asu.edu/courses/122165/files/54912778/download?download_frd=1)**

In this video, Dr. Wilson explains the structure, function, and evolution of the sex chromosomes. Dr. Wilson expl sequence conservation between the X and Y chromosomes makes it difficult to align sequencing reads to the se take-home points that are important for this research experience are to know:

X and Y chromosomes share evolutionary history and were once homologous autosomes.

- The Y chromosome is much shorter than the X chromosome now, but still retains regions of shared ancestry.
- Because of this shared ancestry the modern X and Y chromosomes still share sequence similarity that can cc algorithms. Because of this, we need an approach to correctly represent the sex chromosome complement th samples we are looking at.
- There are regions of the sex chromosomes that are 98 - 100% identical.

Sex chromosome complement is not always XX or XY, and may not match gonads or gender identity.

- Common sex chromosome variations include Kleinfelter syndrome (47,XXY, which occurs in about 1/500 live individuals and Turner syndrome (45,X), which occurs in about 1/2500 live assigned female at birth individual
- Swyer syndrome is uncommon (1/20,000 live births), and results when a person has XY sex chromosomes b

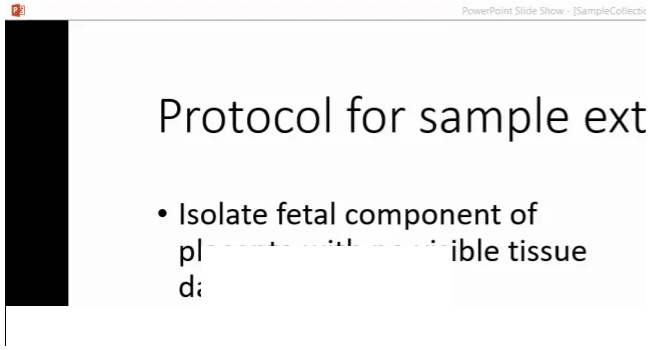
It is possible for the sex chromosomes in a particular tissue to not match the overall karyotpe of the pers

- For example, it is a common phenomenon that an individual who is born XY will lose their Y chromosome wit cells over time as they age.
- Further, it is common for cancer cells to lose an X chromosome (if starting as XX) or a Y chromosome (if start rapidly divide.

The Wilson lab developed a technique called XYAlign which masks parts of the human reference genome to imp the sex of the sample. The XX samples in this study were aligned to the reference genome with the Y chromosor out. The XY samples were aligned to the reference genome with the pseudoautosomal regions which are identica chromosomes masked out, because you can not determine which sex chromosomes those reads came from.

Our Placenta Samples

In this video, Dr. Plaisier discusses how our samples of interest form the placenta were collected following delive pregnancies (> 36 weeks). To minimize confounding factors that could skew results, the patient enrollment criteri women with uncomplicated pregnancies, no delivery complications (e.g. premature delivery), and no medical con hypertension (high blood pressure). The tissue samples were collected from the fetal side of the placenta (indicat cord) and snap-frozen and stored by Yale Biobank. The Wilson Lab purchased these samples from the Yale Biob with the Yale Genome Sequencing Center to extract DNA and RNA and perform whole exome (DNA) and RNA s the .fastq.gz files we will use for this research. We will discuss the fastq files at the end of this module. The next s all members of this class, to process and analyze the data.



Video. Sample Collection.

In this video, Wilson Lab Research Scientist, Dr. Seema Plaisier, discusses the origin of the placenta samples we will be using in this class, from patient criteria to sample extraction and sequencing.

View Transcript (https://canvas.asu.edu/courses/122165/files/54792269?wrap=1) **(https://canvas.asu.edu/courses/122165/files/54792269/download?download_frd=1)**

Genomics and gene expression

What is transcriptomics?

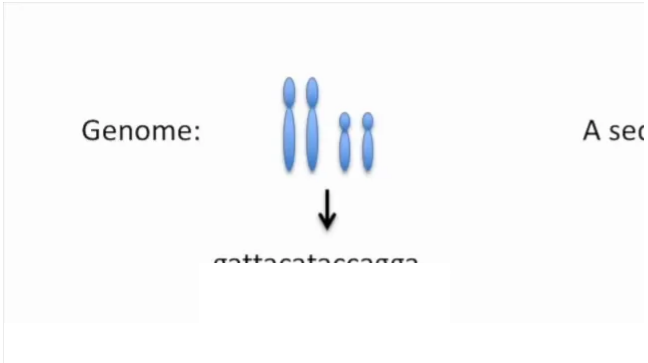
In this course, we will be using next generation sequencing to measure global gene expression, which is one typ (sometimes referred to as transcriptomics). Genomics is the study of the genome (the complete set of genes for those genes interact with one another and the environment. Some of the ways genomics can be used are to prec disease, evaluate drug effectiveness, or to understand biological processes. Transcriptomics is when we assay a (mRNA) in a cell. Using high-throughput sequencing technologies, we can lyse cells from specific tissues, stabiliz

sequencing to detect and quantify the set of mRNA transcripts being transcribed, giving us a window into genes t
specific cellular processes and conditions. We can do this in different tissues in different conditions to tease out
specifically involved in which contexts, thereby allowing us to sleuth out mechanisms underlying the biology of a
A gene is considered expressed when it is actively transcribed; some genes are constantly expressed while othe
expressed, and vary with tissue, age, sex, and environmental response (e.g., stress). If you need more informatic
biology used in bioinformatics, refer to your **Biostar Handbook: Biology for Bioinformaticians (https://www.biostarhandbook.com/biology-for-bioinformaticians.html)** .

RNAseq data pre-processing approaches

In this course we will focus on the RNA that is transcribed in the cells of the human placenta tissue. We extracted
the XX and XY placentas and sequenced that RNA using RNAseq (RNA sequencing).

We have already done the RNAseq data processing for you but it is important as a researcher to understand whe
from and how it was processed. In the video below, we will explore what RNAseq is, the processes behind RNA s



Video. StatQuest: A Gentle Guide to RNAseq.
In this video, you will learn what RNAseq is, how biological samples go from *in vitro* (test tube) to *in silico* (virtual), and what that means for downstream analysis.
View Transcript (https://canvas.asu.edu/courses/122165/files/54792276?wrap=1) ⬇
(https://canvas.asu.edu/courses/122165/files/54792276/download?download_frd=1)

(https://canvas.asu.edu/courses/122165/files/54792276?wrap=1) RNAseq is a type of high throughput or next-gener method that tells us which genes are active by measuring the amount they are transcribed. In the last section, we that are active produce mRNA transcripts while those that are inactive do not, but there could also be differences mRNA transcripts being produced between two conditions. If we take normal cells as our control and compare th differences between mutated cells, we will see that some genes are expressed in both conditions at the same lev amounts of transcripts for those genes while some genes have increased expression in one condition producing transcripts.

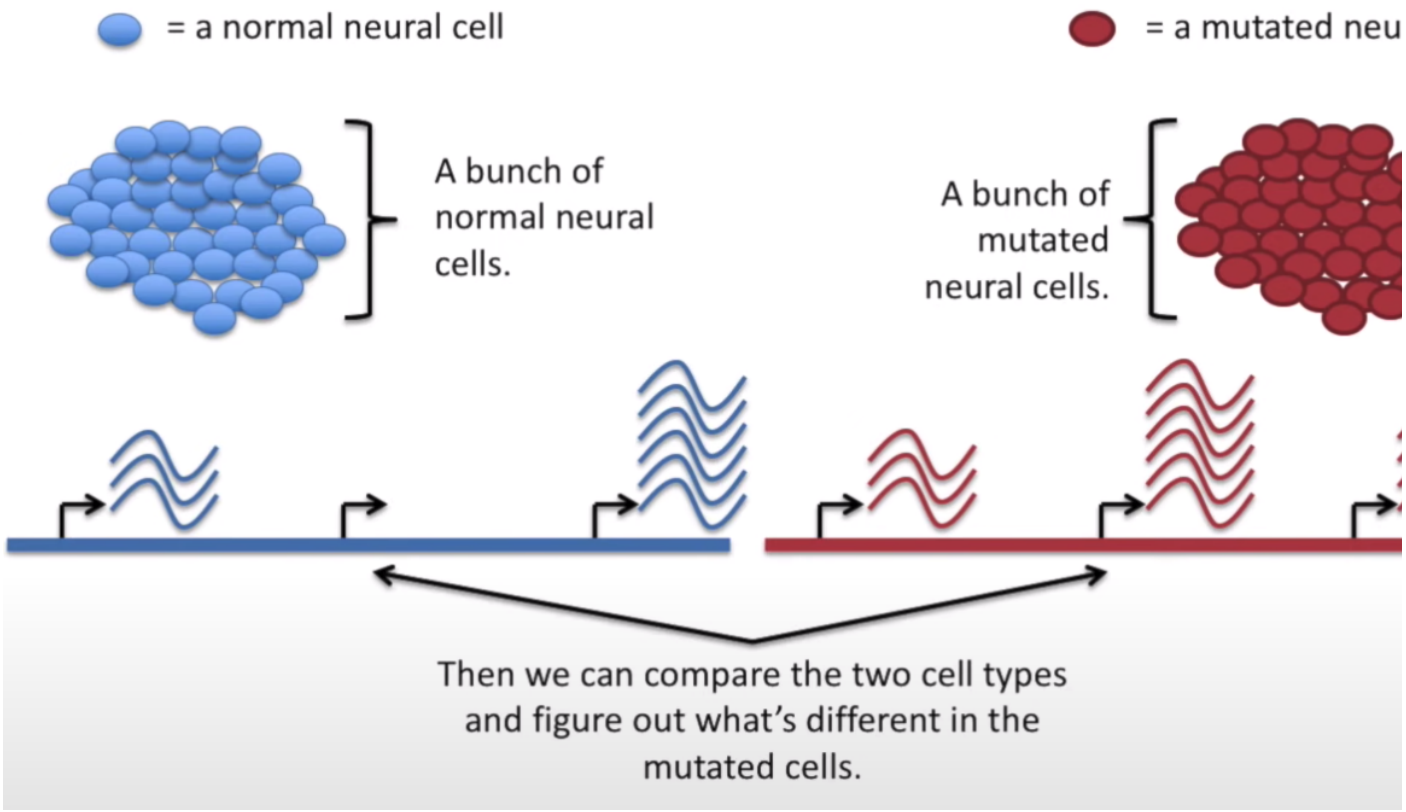


Figure. RNA sequencing overview.
We can use RNAseq from normal cells as our control and compare the results to mutated cells to find differences in gene expression. In the example, each right arrow indicates a transcription start site, or beginning of a gene, and the wavy lines beside them depict the amount of mRNA transcripts being produced by each gene.

There are three main steps for RNAseq: preparing a sequencing library, sequencing, and data analysis. Let's tak each step is performed.

Step 1: Preparing a sequencing library

Prepare a sequencing library (based on the protocol from Illumina, the company who makes the library prep kit w

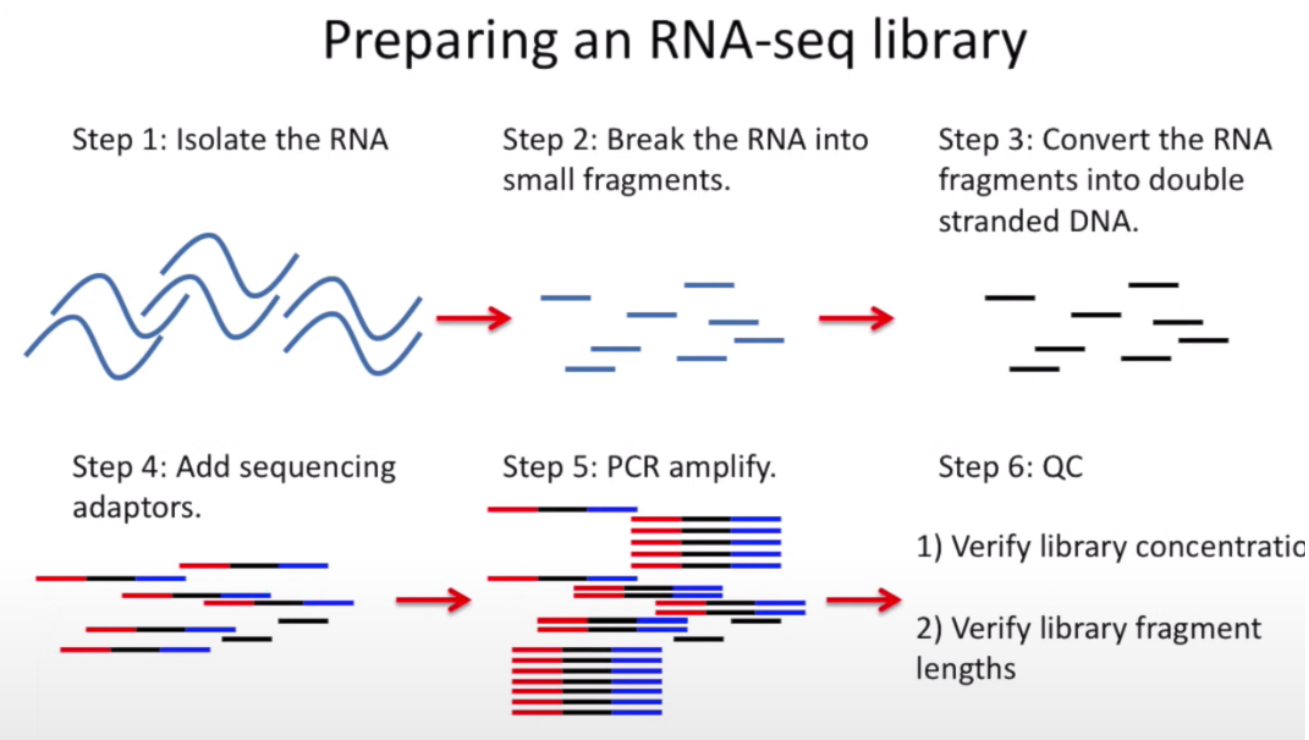


Figure. Preparing an RNAseq library.
The steps in preparing an RNA-seq library are to isolate the RNA, break it into small fragments, convert it into cDNA (complementary DNA), add sequencing adaptors, PCR amplify, and then perform QC to verify library concentration and fragment length.

1. Isolate the RNA from cells that were treated to prevent RNA degradation (stored in cold temperatures with RN Break the RNA into fragments because the sequencing machine performs best with 200-300bp fragments an can be thousands of bases long.
2. Convert RNA into double stranded cDNA because DNA is more stable than RNA and can be amplified using a
3. Add sequencing primers which will allow the sequencer to read the fragments and identify the nucleotide sequ
4. PCR amplify the library of fragments from theadapter to the end of the fragment

5. Perform quality control to verify library concentration and library fragment lengths to ensure the sequencing re
and precise (quality control to make sure there were any unexpected technical problems).

Step 2: Sequencing

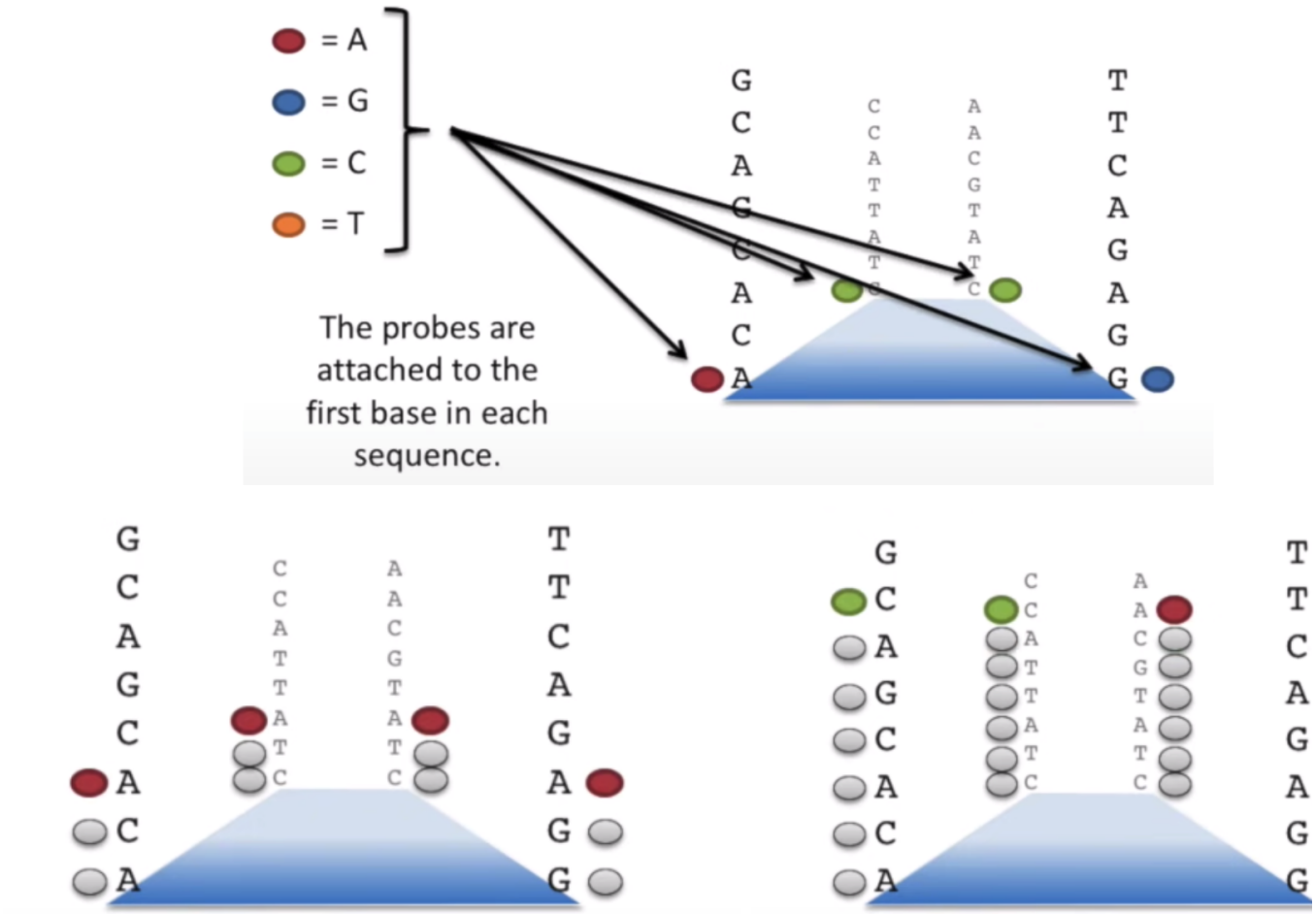


Figure. The process of sequencing using a flo cell and fluorescent probes.
The probes attach to the first base in each sequence, are photographed, and then washed away; this process is repeated until the full sequence of each fragment is determined.

After the library is prepared with fragmented cDNA that has been PCR amplified, the fragments are laid out in a g
The machine attaches fluorescent probes to the bottom base of each sequence with a different color for each bas
machine takes an aerial photo from above where it can detect each base by color; after it registers the lit base in
washes off the color from the probe. Now the machine can attach the next row of probes and repeat the process
sequence of each fragment.

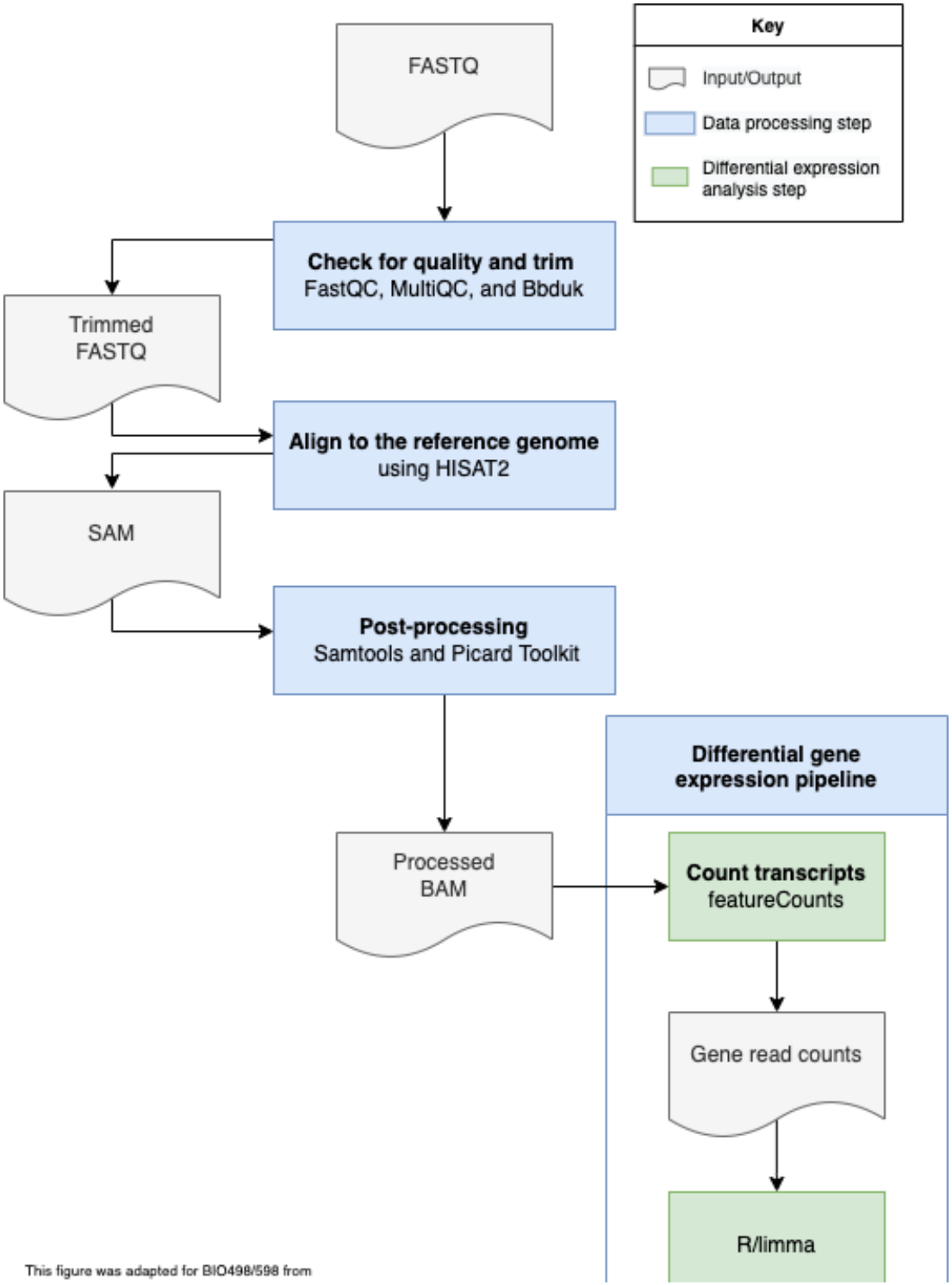
Since the average run can be around 400,000,000 reads, probe colors can be too light to detect or have “low div
probe of one color in a region saturated with probes of another color, making it difficult to identify with confidence
that position. Either of these would be considered low quality bases and, if many are in the same sequence, can
we trim.

Step 3: Data Analysis

After reading the sequences from these fragments, we are given a file that contains the nucleotide sequences for
a huge number of software packages designed to help you make sense of these large files of read sequences. T
software to match these reads to genes is called **data preprocessing**. Using software to compare the amount o
in one set of samples versus another is called differential gene expression analysis.

Data Preprocessing

This diagram shows the typical steps of sequencing data preprocessing along with the specific software we used
Starting from the read sequence files from RNA sequencing (.fastq), we first used tools like FastQC and MultiQC
sequence quality checking and the trimming tool bbduk to trim low quality sequence and adapter sequence. Nex
outliers, the reads are aligned to the reference genome so we know what genes those reads map to. Once align
do some post-processing analysis to make the alignment output files easier to work with (sorting, indexing, etc).
reads mapped to each gene are tallied and used as a measure of gene expression. We will go into more detail w
steps, specifically as they relate to our research aims.



This figure was adapted for BIO498/598 from

Figure. RNAseq data preprocessing and differential expression workflow.
This workflow was adapted for this course from **https://github.com/SexChrLab/Nasonia (https://github.com**
The
lab followed the data generation protocol (left) of checking the sequencing FASTQs for quality, trimming, aligning to a reference genome, and post-processing to generate gene counts. You will conduct the differential gene expression analysis (boxed in blue on the right) of taking the counts and processing them through the R/limma-voom pipeline.

Step 1: FastQC and Trimming

FastQC is a tool that provides data visualizations that can help you to identify samples that had technical issues g of your overall data quality. This step is done routinely and can be referred to at any point in the study if we suspe problems. Trimming software is often used to trim or filter sequencing reads with poor quality. We will use the trir (**https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/bbduk-guide/ (https://jgi.d tools/software-tools/bbtools/bb-tools-user-guide/bbduk-guide/)**) which we have been able to use to improve data q studies with many different types of sequencing data.

The first figure FastQC gives you shows the quality scores across your reads. This is an example of this figure fr set. You can see as you go towards the end of the reads, the quality scores fall to zero, going from a warning lev poor level marked red. Trimming software can be used to filter poor quality sequence reads, reads where the se of which base was at a specific position is not that confident. Results as bad as what is shown in this example w completely with trimming software, but most data with a smaller proportion of poor quality reads can be filtered su steps such as alignment to the reference genome proceed more smoothly. The parameter 'trimq' in bbduk is use quality score. If bases in a read are less than the minimum threshold, they are trimmed off in the output fastq se parameter of bbduk called 'minlen' sets a minimum length of read that has to be left after quality trimming. In the thumb is that if trimming removes over half of the read, then we should filter out that read as a low quality read. S will be using, the reads were 150 base pairs in length, so the original value of minlen we chose for data processir investigating the effect of changing those values in this research project.

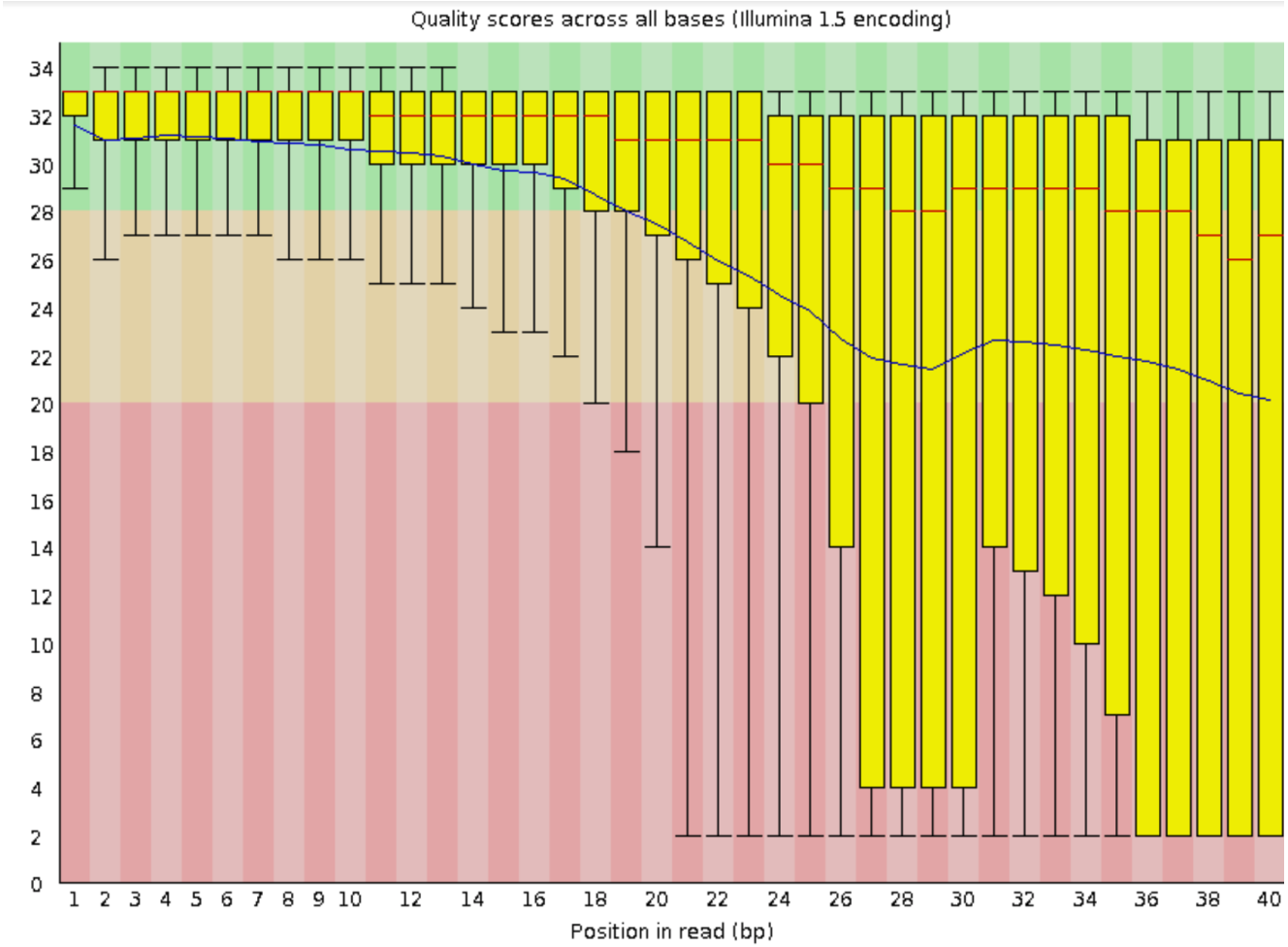


Figure. Quality scores from poor quality RNAseq data.
The boxplot was generated using FastQC and shows the quality scores across all bases using Illumina 1.5 encoding. The x-axis is the position of the read in basepairs (bp) and the y-axis shows the quality score (PHRED).

Another FastQC figure that is relevant to our research shows the percentage of Illumina sequencing adapters. Th
the difference in adapter content before and after trimming with bbduk. Each line on the figure represents one sa
The x-axis of the figures show the position along the 100-base pair reads and the y-axis shows the proportion of
to an Illumina adapter used during library preparation at that position. We should not see Illumina adapter seque
because they are supposed to be removed as part of the Illumina library generation protocol before the sequence

(.fastq). This is reflected in the max value of the y-axis is 7%. Since the library prep kit adapter removal step is n in the untrimmed data figure on the left that the overall percent adapter sequence is low, but some samples have expected for high quality data (ranges marked in yellow for warning, marked in red for poor data quality). Remov sequence is one of the key functions of trimming software and improves alignment of these reads to the referenc see that after computationally trimming with bbduk, the percent adapter sequence falls below 5 percent for all sar analysis, any data set for which trimming was done at all included removal of Illumina adapter sequences, regard trimq and minlen.

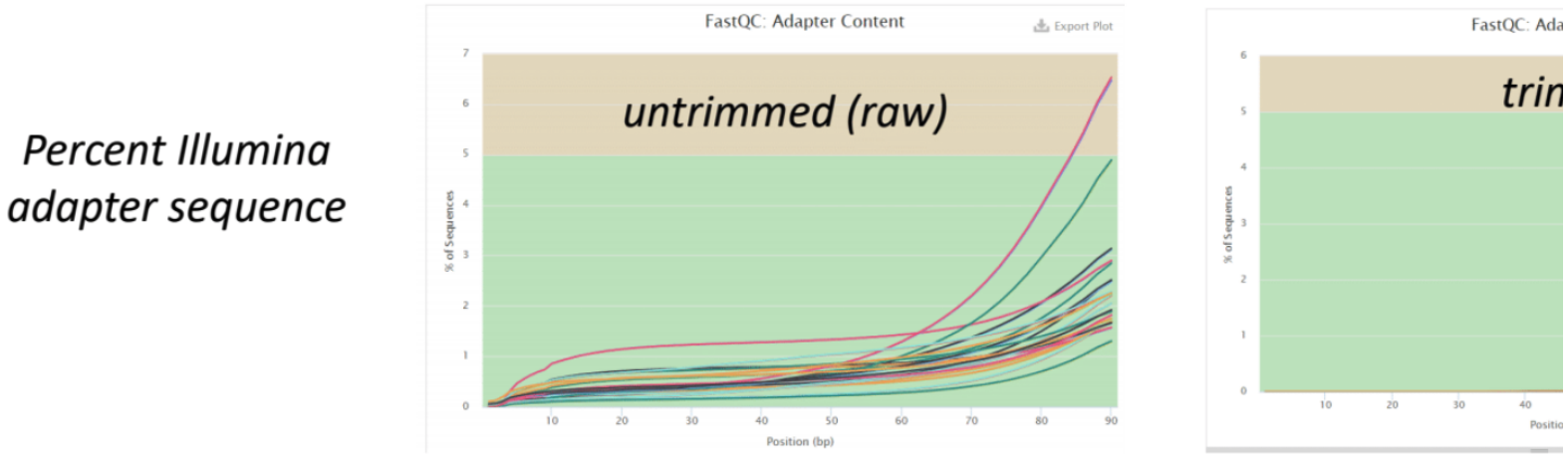


Figure. FastQC figures of percent Illumina adapter sequences.
We can use this to compare adapter content from our untrimmed batch to the trimmed batch.

Step 2: Alignment

To perform alignment, we can download the sequence of the entire genome for the species of the organism whos analyzing, human in our case. We used the GRCh38 build of the human genome, an established human genom of publications. Since we are using sequencing at the transcriptome level, we downloaded the entire set of transc genome build to use as an alignment reference. Software to align short read sequences to the reference genom we used the HISAT2 algorithm to do this step.

Aligning to autosomes (non sex chromosomes) is generally quite straightforward as genes can be expressed fror copies of each autosome which are nearly identical. Alignment to the sex chromosomes is more complicated be

chromosomes are quite different from one another in size and sequence, but contain specific regions of identity d
evolutionary history. We need to make additional considerations in order to accurately align reads to the sex chro

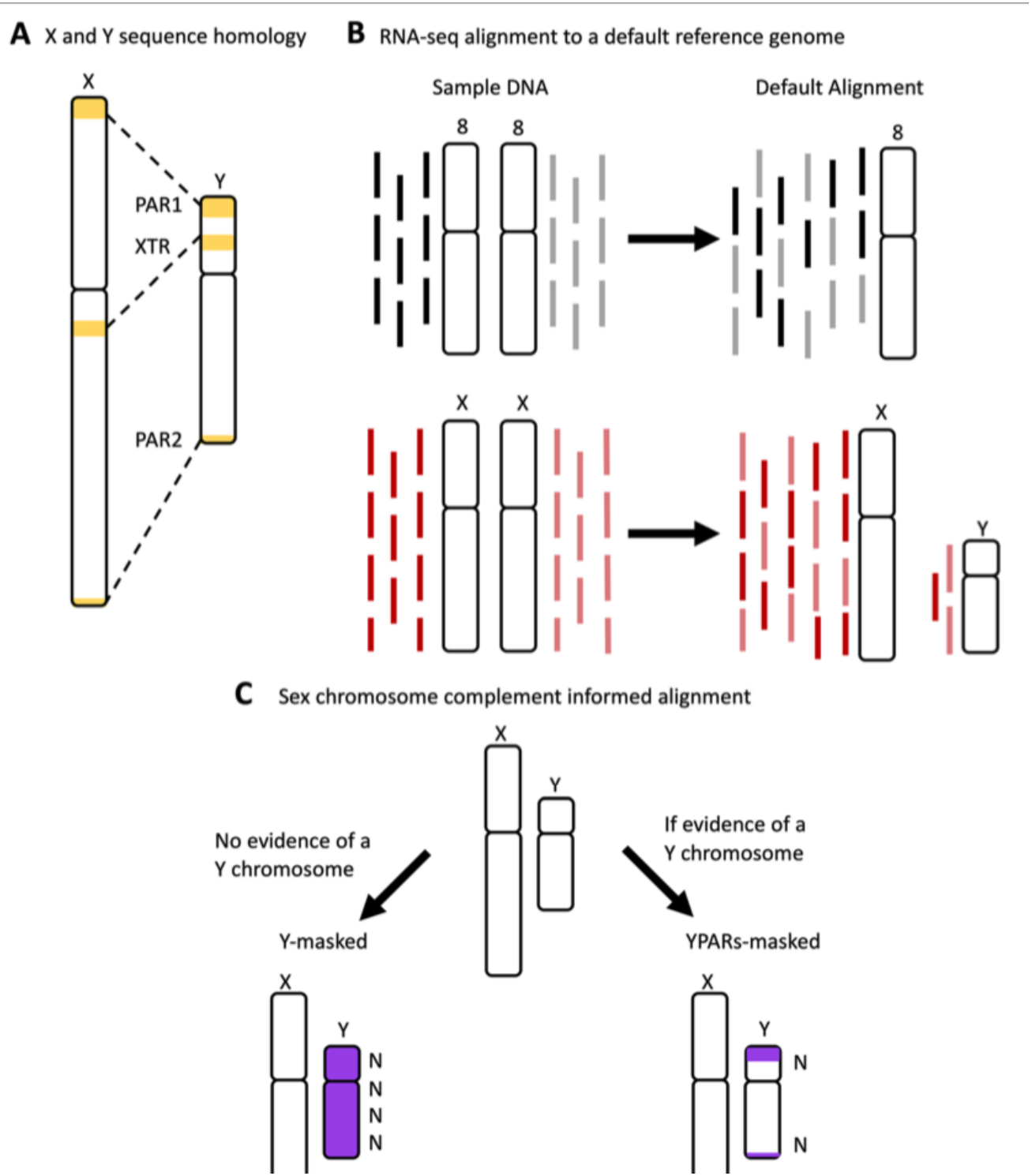

A X and Y sequence homology
X
Y
PAR1
XTR
PAR2
B RNA-seq alignment to a default reference genome
Sample DNA
Default Alignment
8
8
8
X
X
X
Y
C Sex chromosome complement informed alignment
X
Y
No evidence of a
Y chromosome
If evidence of a
Y chromosome
Y-masked
YPARs-masked
X
Y
N
N
N
N
X
Y
N
N

Figure. Sex-informed alignments to a reference genome.
(A) X and Y sequence homology, (B) RNAseq default alignment to reference genome, (C) sex chromosome complement informed alignment.

This figure depicts how we do alignments taking sex into account. Figure (A) illustrates the sequence homology chromosomes, XTR and PAR regions. In (B), the standard RNAseq alignment shows that the sample DNA from b chromosome 8 align to the sequence of chromosome 8 in the reference genome. This works for autosomes, but the sex chromosomes? Since there is sequence for both the X and Y chromosomes in the reference genome sec download, so reads from the homologous XTR or PAR regions from an XX sample could randomly align to chrom the sample has no chromosome Y. Similarly for an XY individual, reads that are mapping to the regions found in b chromosomes will be randomly aligned to either the X or Y chromosome sequence in the reference genome. Thi samples that have a Y chromosome differently from samples that do not have a Y chromosome.

The solution (and way we processed our dataset) is depicted in figure (C), using a sex complement informed alig technique called masking, which replaces specific sequences in the reference genome with “N”s, such that no se those masked areas and those reads are forced to align somewhere else in the genome. To prevent shared seq misaligning, we prepare a reference genome with the Y chromosome sequence masked to use for alignment of r samples. We prepare a second reference genome sequence file with the PARs masked for alignment of XY samp not really know where those reads are supposed to map. Using this alignment method, we were able to see diffe of many genes on the sex chromosomes and characterize those genes in a manuscript.

Step 3: Gene Expression Quantification

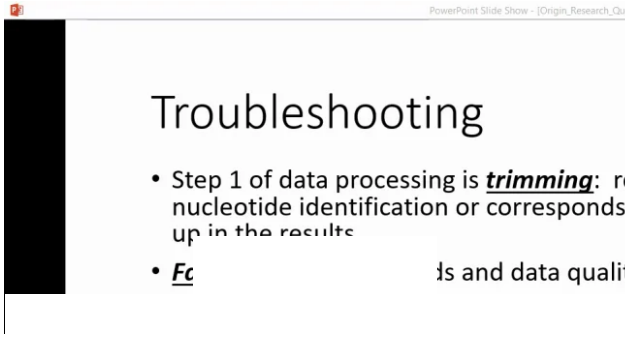
After alignment, the number of reads that map to a specific gene is used as a way to quantify gene expression, s align our reads, the better our estimation of gene expression and the more likely we are to pick up true difference for genes on the sex chromosomes. We use a program called featureCounts to take the coordinates of genes giv file we downloaded from the same database as the reference genome sequence, featureCounts uses the output to count the number of reads that map to each gene. These are given as a single file per samples. We use a sir merge all of those files into one table that contains the counts for each gene in all samples in our study.

Differential Expression (DE) Analysis

You will learn a lot more about the differential gene expression pipeline we will be using later in the course, but w
broad overview here. Our aim after data processing is to find genes that are truly expressed at a different level ir
placentas. To do this, we first filter out genes with no reads detected in our samples. We next normalize the gen
account for technical issues such as the fact that longer genes are going to have more reads coming from them e
expressed at a higher level. Once the data has been normalized, we can use data visualization techniques to gr
to their gene expression profiles to see if they cluster by sex as well as use molecular modeling to find genes are
significant differences in expression between XX female and XY male placentas.

How we came up with the research question for this course

Now that you have a good idea of the biological system we are studying and the way that we are measuring gene
understand that system, we wanted to tell you how the idea for this project came to be. As with all scientific rese
an observation. In the video below, Dr. Seema Plaisier explains how applying an established data processing wo
of samples led to the research question we are interested in. She discusses how our lab applied trimming param
the placenta RNAseq we will be analyzing in this course to a new set of placenta RNAseq. In the original researc
parameter values we used (trimq=30, minlen=75) gave us data that passed all quality control requirements and c
loss in overall read count after trimming. However, in the sequencing data for the new placenta samples, trimmir
parameters resulted in overfiltering and erroneous results. This got us thinking about how trimming affects downs
class, we will be processing the samples from the original study using a range of values for trimming parameters
investigating what changes in the lists of genes differentially expressed in XX vs XY placentas.



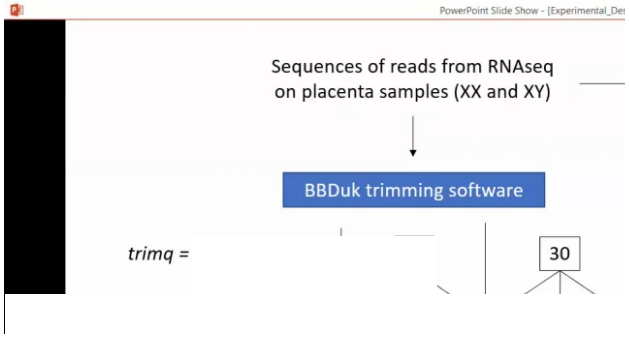
Video. Origin of the research question.

In this video, Wilson Lab Research Scientist, Dr. Seema Plaisier, discusses the origin of the research question and how it was organically derived from replicating research workflows.

View transcript. (https://canvas.asu.edu/courses/122165/files/54792240?wrap=1) **(https://canvas.asu.edu/courses/122165/files/54792240/download?download_frd=1)**

As explained in the video above, the main underlying question we are asking in this course is: ***How does trimmi gene expression analysis?***

Since we were questioning the trimming parameters of the original study, we wanted to get an idea of what other did a quick search for published research papers that used the same trimming software (BBDuk) to trim RNAseq samples. We found that of all the trimming parameters, trimq (trimming quality) and minlen (minimum length of re quality trimming) were the most frequently reported in the methods and the most highly variable. We used this to that spanned the range observed in literature and seemed appropriate given our experience with using BBDuk in projects.



Video. Experimental design.

In this video, Wilson Lab Research Scientist, Dr. Seema Plaisier, discusses the experimental d
of our research, the trimming parameters we are investigating, and why.

View transcript. (https://canvas.asu.edu/courses/122165/files/54792262?wrap=1)

(https://canvas.asu.edu/courses/122165/files/54792262/download?download_frd=1)

For this course we chose to investigate a range of 0, 10, and 30 for trimq and 10, 30, and 75 for minlen to test for
on differential gene expression analysis. We used a workflow management tool called Snakemake to generate t
use for this study in an automated, reproducible way. The datasets we will be working with consist of one set of u
reads and nine sets of trimmed datasets. For the untrimmed reads, we did not apply any trimming, therefore, do
minlen parameters. The untrimmed dataset will serve as a control and the entire class will be using the same cor
We generated nine trimmed datasets by running the untrimmed data with bbduk using all combinations of trimq (
minlen (10, 30, and 75). The untrimmed and trimmed data sets were independently run through data preprocessi
putting them through differentially gene expression analysis so that the results can be examined to find trends. E
analyze one trimmed dataset then we will compare results in groups and as a class to analyze the downstream e
trimming parameters.

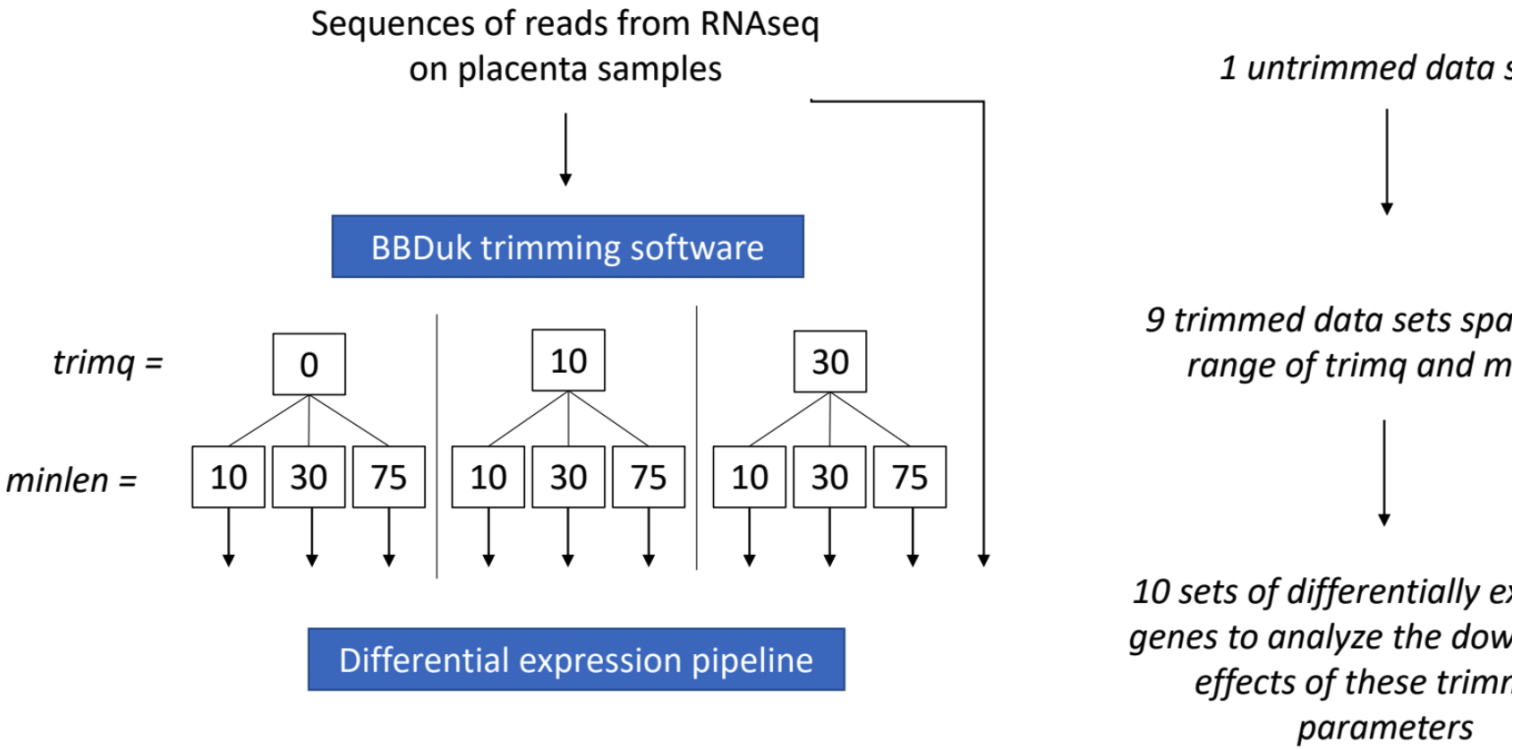


Figure. Experimental design of trimming parameters.
We will be using one (1) untrimmed dataset and nine (9) trimmed datasets with a range of parameters to generate ten (10) sets of differentially expressed genes to analyze.

In the figure above, you can see the nine trimmed datasets that have a different value for trimq and minlen param generated for you. Each person has been assigned the control plus one of the following:

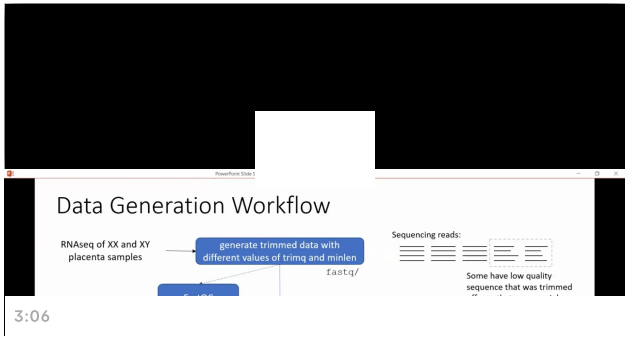
Table. List of the trimming parameter combinations selected for this experiment.

Directory name	Trimq	Minlen

`trimq_NONE_minlen_NONE`	NONE (control)	NONE (control)
`trimq_0_minlen_10`	0	10
`trimq_0_minlen_30`	0	30
`trimq_0_minlen_75`	0	75
`trimq_10_minlen_10`	10	10
`trimq_10_minlen_30`	10	30
`trimq_10_minlen_75`	10	75
`trimq_30_minlen_10`	30	10
`trimq_30_minlen_30`	30	30
`trimq_30_minlen_75`	30	75

Data Generation Workflow

This video will provide a quick recap of the data preprocessing steps to create the gene count matrices you will b
class.



Video. Data generation workflow.

In this video, Wilson Lab Research Scientist, Dr. Seema Plaisier, discusses the data generation workflow that has already been conducted on the dataset you will be using throughout the course. RNAseq samples were trimmed with the different parameters, evaluated for quality, aligned to a sex-informed reference genome, and then quantified into gene counts.

View transcript. (https://canvas.asu.edu/courses/122165/files/54792240?wrap=1) ⬇ **(https://canvas.asu.edu/courses/122165/files/54792240/download?download_frd=1)**

Data Generation Workflow

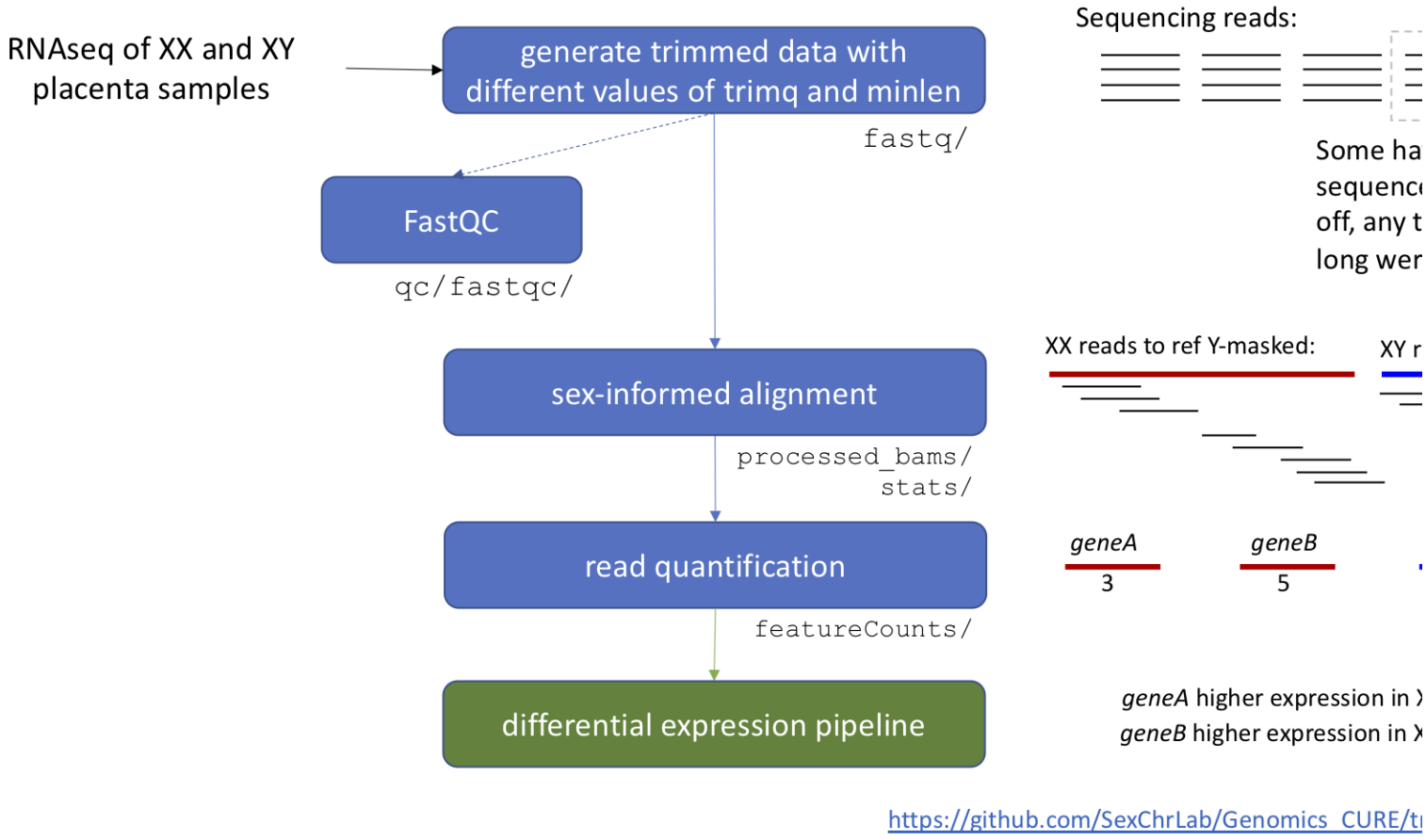


Figure. Data Generation Workflow.
The diagram on the left shows the steps that have already been conducted for you (blue) and the analysis you will be conducting (green). The figures on the right are examples of the corresponding step to the left.

Module 1.1 Additional Resources

- **Genomics** (https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics)
- (https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics) Placenta (detailed article Burton GJ, Fowden AL. The placenta: a multifaceted, transient organ. Philos Trans R Soc Lond B Biol Sci. 20 5;370(1663):20140066. doi: 10.1098/rstb.2014.0066. PMID: 25602070; PMCID: PMC4305167. **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4305167/** (https://www.ncbi.nlm.nih.gov/pmc/articles/PM
- **Central Dogma** (https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/) :
- **Biostar: Biology for Bioinformaticians** (https://www.biostarhandbook.com/biology-for-bioinformaticians.h
- **Snakemake** (https://snakemake.readthedocs.io/en/stable/tutorial/basics.html)
- The X Chromosome **TED-ed: secrets of the x chr [5:05]** (https://youtu.be/veB31XmUQm8)
- Clifton. (2010). Review: Sex and the Human Placenta: Mediating Differential Strategies of Fetal Growth and S (Eastbourne), 31(3), S33–S39. **https://doi.org/10.1016/j.placenta.2009.11.010** (https://doi.org/10.1016/j.pl
- Olney, Brotman, S. M., Andrews, J. P., Valverde-Vesling, V. A., & Wilson, M. A. (2020). Reference genome an informed by the sex chromosome complement of the sample increase ability to detect sex differences in gene RNA-Seq data. Biology of Sex Differences, 11(1), 1–42. **https://doi.org/10.1186/s13293-020-00312-9** (https://doi.org/10.1186/s13293-020-00312-9)