Administrivia:

- 4-5 people per group
- Homework: look into project ideas, potentially think of your own if you're not into the ones Ruzzo suggested
- Biggest challenge is management
    - Each group will produce a summary for each week and a schedule of what to do the week after (status + projection)
- End of quarter will have written and oral presentations
- Class Sessions will be the groups working, but you still need to have a comprehensive working schedule
- Should have some amount of bio + CS
- Projects are relatively open ended

Background:

- Cells have DNA (23 pairs of chromosomes)
    - Modern sequencing technology can read this DNA in terms of bases (1% error sometimes)
    - Error rate is relatively easily managed with "majority rules" approximation
- Sequencing the Human Genome is a stepping stone to do other stuff
- DNA is a template to make RNAs (which make proteins and other stuff)
    - DNA is static in most cells
    - Biochemical protocols allow us to sequence DNAs (map them back to the genome to figure out where they came from)
        - Get the RNA out of a cell, get the DNA, fragment the DNA, map them to the genome
    - Different genes do different things in different parts of your body : largely represented in the RNA sequences and proteins developed
    - Genome sequence is the means to an end, not the end itself

Bias in RNASeq data:

- Which genes are being expressed? How highly expressed are they? What's the same/diff between 2 samples?
- Genes are often interrupted by parts of the genome that don't affect proteins (introns)
- We expect uniform sampling across a gene for reads (but it's actually highly non-uniform)
    - Can we make it more uniform? Averaging/smoothing
    - It's better to model the aspects of causation
- Bias is sequence dependent:
    - It's easier to capture reads that start with an A than with an C
        - Why is that? It's a technology problem
        - Based on data modeled with Illumina
    - Model the bias, it'll be useful to understand dependencies later
        - Need to estimate 32,000 parameters (use Machine Learning)

- Daniels Method:
    - Sample foreground, sample background (based on where the read came from)
    - Train a Machine Learning Model (Bayesian Network) to predict the bias:
        - For every sequence of 40 bases, what will the bias be?
    - Directed Bayes nets model:
        - Look at a 41 base window, filled nodes mean that the position is bias, and the arrows mean that the first node modifies the bias of the second node
        - 5 datasets show that they all have different biases
    - Is the problem still this bad with modern data? How much does it vary from one dataset to another?
    - More training data allows you to get a better fit (flattens out a bit), but training time increases
    - What if the input wasn't biased?
        - Does this make the data look biased?
        - Nope, not with this method, the probability actually goes down
    - How much of the difference is a result of biology vs bias?
- Batch effects for determining correlation
    - The software matters in determining this correlation, bias correction is important here

Project Ideas:

1. Bias Visualization and Exploration in datasets
    - ○ Build a tool to do this
    - ○ Explore RNASeq Datasets
    - ○ What is state of the art right now for examining an RNAseq dataset?
    - ○ Visualization issues
    - ○ Add SeqBias, other visualizations, look at modern data
    - ○ Time permitting, explore how to find the cause of bias in this data
2. Allele Specific Expression
    - ○ You have two different alleles but you only express one of them
    - ○ Not super understood how widespread the phenomenon is (likely a grayscale and not binary)
    - ○ How do we recognize Allele Specific Expression? Potentially use RNASeq experiment, recognize the genes that account for alleles
    - ○ Biologically interesting: Male/Female competition in reproduction
    - ○ What is the effect of bias in RNASeq here?
    - ○ What is the State of the Art?
3. Determine RNA secondary structure
    - ○ Are these techniques affected by RNASeq bias?
4. Next time

Homework:

- Skim the paper assigned for Thursday
- Skim Allele Specific Expression paper