Organization
- Group Projects
- 19 people, 4-5 groups (4-5 people, maybe smaller)
- pick your own partners
- Class sessions are groups working, pretty much
- Going to try to move to a class room w/movable desks
- Class is grounded in biology, but not deeply dependent
- The point is to get experience with management, tools, groupwork, etc.
  - There are opportunities for algorithm development and statistical analysis of data, coding experience, HCI experience
  - Opportunity to use tools from different areas

Homework:
- Reading to find supporting info on projects
- More detail next lecture
- Skim Daniel’s paper for Thursday
- Skim an ASE paper
- Skim an RSE seq quality control paper

Projects:
- Can come up with your own – need some biology, and some computer
- Can all work on same project, doesn’t matter
- Most projects involve finding an existing tool and understanding it thoroughly
- Ideas:
  - Genome assembly (see website)
  - Bias in RNA Sequencing
    - #1: Bias Viz + Explore – we need a convenient tool for this
      - What is state of the art?
      - What’s out there for quality control in RNA datasets?
      - What’s normal?
      - Can we graft something smoothly onto the open source tool to present this bias?
**Exploration** – is it better with modern data? How much variation Monday to Tuesday, lab to lab, etc.?

One challenge is reducing this data to be understandable

Other visualizations...?

- **#2: Allele (version of gene) Specific Expressions (ASE)**
  - You have two alleles, but you’re only expressing one of them
  - You may only express Mom or Dad’s version of gene
  - Not super well understood
  - Likely a gray-scale (60% Mom, 40% Dad or something)
  - Just a few differences in versions
  - You could look at the RNA sequencing and see if you have a lot more of A allele than B allele then maybe that’s important
  - But what about bias?
  - What’s the state of the art for detecting ASE from RNA seq?
    - Note that genome sequencing gets you one copy, not two, so one allele is in the reference for sequencing
    - It’s a mosaic of individuals, not one person
  - Does bias correction help?
  - Can you quantify that?

- **#3: RNA Structure and Seq Bias**

- **#4: TBD**

**Grading:**

- The challenge is management (and jargon)
- Each group each week produces a report:
  - what we did last week,
  - what we will do next week,
  - long term/medium term goals,
  - how did we do on last weeks goals,
how we subdivided work

- End of quarter written and oral presentation
- Everybody will get an A – that’s the plan (usually works out)
  - Don’t worry if everything doesn’t work out perfectly

Background Bio:

- Cell DNA – 23 chromosomes
- Modern sequencing tech can decode sequences of chunks of DNA very, very efficiently
  - 1% mislabeled due to glitches
  - so do it with redundancy
  - sequence the tiny chunks and you can piece it together with reasonable confidence
  - error rate can be managed via majority rules
- But sequencing is just a stepping stone
- Much of DNA is a template for making RNA
- We can extract RNA, convert to DNA, and sequence it like you would the genome
  - You can then map it back to genome to find where they came from
  - Or just treat it like usual DNA and sequence it
- All cells have the same genes, but the genes are obviously doing different things in different cells
  - That’s mostly down to RNA
- Genome has been annotated to show different genes (partially via sequencing RNA and mapping it back)
  - Depending on where the RNA is from you can figure out where that gene is active (nose, liver, etc.)
  - The gene is “expressed” (making RNA, protein, etc.)
  - More active genes may be more important

- **Bias in RNA sequencing:**
  - Potential questions related to RNA sequencing:
    - Which genes are expressed
    - How much
    - What’s same/diff between two samples (e.g. tumor v normal)
Genes can be ‘interrupted’ by DNA that is irrelevant to the ultimate RNA or protein that’s produced

Remember, this is highly simplified

From this naive model (DNA \(\rightarrow\) sequence \(\rightarrow\) count) we expect uniform sampling across the starting points

But we don’t get that!
  - It’s highly variable
  - The peaks are huge!
  - You can use math tricks to smooth it a bit
  - But we want to know why – what’s causing this?

One of the TAs has a program that helps quite a bit (38% reduction in LLR)

Something about the tech makes it easier to capture some fragments than others (no idea why!!)
  - We’re somehow introducing bias along the way
  - (on graphs) prefer C and G to start, but A and T for second position
  - Something about the sequence effects the success of the capture

One approach:
  - Correct for the most common 7-letter sequence starts
  - We need to look at what is to be expected for expressed genes
  - We have so much data, that 32000 free parameters isn’t really a problem

Method outline
  - Sample foreground (what we get from sequencing)
  - Sample local background (actually sample the expressed genes)
  - Train Bayesian network
  - Predict bias
  - Adjust read counts

Form of the Models:
  - Directed Bayes Nets
    - Black dots mean position contributes to bias
    - +/- 20 bp from start of read
- arrow shows what positions affect other positions
- But all the datasets produce different nets!
  - And we have enough data to be confident of them
- They all get positions outside the read itself affecting the read, too

  - A few questions:
    - **This data is old, how does new tech affect things?**
    - **How much variation between data sets?**

  - More data means less likely to falsely infer bias, more accuracy, and more runtime (10,000 – 50,000 is a good starting point for training data)

- **Batch Effects**
  - There can be correlation between samples
  - There are some very obvious patterns
  - Software makes a difference here, you get very different plots
  - In these graphs, these are all the same samples from the same place, so they should be highly correlated!
  - Daniel’s is in the top left (pretty good)
  - But many don’t attempt bias correction, and it doesn’t look great