1. **Next-gen sequencing**

   1. You have found a new strain of yeast that makes fantastic wine. You’d like to sequence this strain to ascertain the differences from *S. cerevisiae*. To accurately call a base pair, you need at least 4x coverage and the yeast genome is 12 million base pairs. You submit your sample for a spike-in run and you get $2e5$ 2x150 reads back. Will you be able to accurately call 95% of the genome? What fraction will you be able to accurately call?

2. You read a paper about lactose intolerance and found that a regulatory region in the intron of MCM6 controls the lactase (LCT) gene. Adults that are homozygous for a common variant at this allele are lactose intolerant (let’s call the lactose intolerant allele “l” and the tolerant allele “L”). You would like to know if you are lactose intolerant, so you PCR up the locus and sequence the DNA on a spike-in run. Unfortunately, you only get 5 reads back. All 5 reads contain the “l” intolerance variant. Ignoring sequencing errors, what is the probability of observing 5/5 l allele reads if you are heterozygous (Ll) at that locus? Given that information, should you stop drinking milk?

3. Recently, Oxford nanopore sequenced a human genome at 99.8% accuracy. Is this useful? If not, why? Your friend claims by obtaining more and more sequence coverage, the assembled sequence can be made arbitrarily accurate. Under what assumptions is this true? Do these assumptions hold for Oxford nanopore data?
2. Homology

If you looked at the sample questions, you would know that last year we sequenced a half genome of the alien species DT by obtaining a piece of its degenerated brain. You searched through its genome and predicted a mysterious peptide sequence that might explain how DT hijacked the US election.

You run BLAST using the peptide sequence below, to look for potential orthologous human proteins.

**H M N K E S P R K L A G A G L A**

Given The default parameters for Blast use the BLOSUM62 substitution matrix (shown below), a gap opening penalty of –11, and a gap extension penalty of -1.

|   | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
| A | 4 | -1 | -2 | -2 | 0 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -2 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 |
| E | -1 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | -3 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | 0 | -2 | 3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | -2 | 1 | 3 | -1 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -3 | -1 | -3 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | 1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | 3 | -2 | -2 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | -2 | 0 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -1 | -3 | -1 | 1 | 4 | -3 | -2 | 11 | 2 | -3 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -3 | -2 | -2 | 2 | 7 | -1 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -3 | -3 | 3 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 | -1 |

You obtained one hit which can be aligned in two different ways:

**H M N K E S P R K L A G A G L A**

**H M O R D T P K — — — A G L A**

**H M N K E S P R K L A G A G L A**

**H M O R D T P K — A G — — L A**

1. (1 pt) What is the percentage identity for each pair of alignments?

2. (2 pts) Which is the better alignment and why?
3 (2 pts) Please calculate the P-value of getting a score greater than or equal to the scores from the better of the two alignments, assuming EVD (extreme value distribution): \( P \)
use \( \lambda = 0.693, \mu = 3. 
\[
P(S \geq x) = 1 - e^{-\frac{\lambda(x-\mu)}{\lambda}}
\]
Based on the p-value, do you think you found an ortholog?

If we compare the BLOSUM 45, BLOSUM 62, and BLOSUM 80 matrices we find the following:

In BLOSUM 45 an A-A alignment has a score of 5, and a W-W alignment has score 15;
In BLOSUM 62 an A-A alignment has a score of 4, and a W-W alignment has score 11;
In BLOSUM 80, an A-A alignment has a score of 5, and a W-W alignment has score 11.

4. (2 pts) Why is W-W always higher than A-A?

5. (2 pts) Why is W-W in BLOSUM 45 higher than in BLOSUM 80?

6. (1 pt) Do you think using BLOSUM62 is a good choice? Should we use BLOSUM45 or 80 instead? Why?
3. Expression

a) What is the False Discovery Rate (1pt)?

b) If you are testing 1000 null hypotheses and determine that you can reject the null hypothesis in 40 instances, how many false positives do you expect at an \( \alpha = 0.05 \) (1pt)?

c) If you are testing 1000 null hypotheses and \( \alpha = 0.05 \) is your desired Family-Wise Error Rate, what is the Bonferroni Corrected p-value (1pt)?

d) What are three things that increase your statistical power to detect differential expression in an RNAseq study (1pt)?

e) What is a phiX spike-in and how is it useful (1pt)?
f) Why would degraded RNA result in a 3' bias (2pt)?

g) What are two reasons you would discard a read in a transcriptomics study (1pt)?

h) What does the RPKM/FPKM normalization correct for? What does the TMM normalization correct for? Which is more appropriate for estimating differential expression (2pt)?
4. Epigenetics

In mammals, DNA methylation occurs mainly at CpG dinucleotides. Methylation of the promoter suppresses gene expression, but functional role of gene-body DNA methylation has yet to be determined. You decide to investigate a potential relationship between DNA methylation and histone modification in gene bodies. You took mouse embryonic stem cells and performed whole genome bisulfite sequencing (WGBS), RNA-seq, ChIP-seq on Dnmt3b (a de novo DNA methyl transferase), and ChIP-seq on H3K36me3.

Questions (10pts)

4.1. (2 pts) Describe why we can use bisulfite sequencing to detect DNA methylation.

4.2. (2 pts) Given a double stranded genomic region AC\textsuperscript{m}GTTCGCTTGAG, what do bisulfite reads that fully cover it look like?

You decided to examine epigenetic pattern as a function of gene expression. You ranked all genes based on their expression level from lowest to highest, then normalized genes such that they all have the same length, starting from TSS, and ending at TES, and generated a heatmap of epigenetic signatures (show in figure below). You groups genes into 4 classes of based on their expression level, lowest 25%, 25%-50%, 50%-75%, and top 25%, and generated meta gene plot as shown in figure below. The figure included ChIP-seq of Dnmt3b and H3K36m3.

4.3. (2 pts) Complete the figure by adding DNA methylation metagene plot.

4.4. (2 pts) Describe what you can learn from the Figure. You need to address the observed dynamics and correlation between different epigenetic marks.
4.5. (2 pts) Describe a model that is consistent with the data, and describe one experiment to test the model.
5. DNA binding specificity

In what will later be called the discovery of the century, you have identified a class of DNA binding factor modules called butter finger, as apposed to zinc finger. Given a special type of lipid co-factor, these fingers have extremely strong specificity, such that there are four fingers each recognizes one DNA base. This means:

<table>
<thead>
<tr>
<th>Frequency of:</th>
<th>FingerA</th>
<th>FingerC</th>
<th>FingerG</th>
<th>FingerT</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

Because the fingers are so modular, you can design any combinations of fingers and make a TF that recognizes a specific sequence! Forget about CRISPR! It’s all about butterfinger and you now!

1. (3 pts) How many fingers must be combined to achieve an information content of 20 bits in a genome with uniform background distribution for the four bases?

2. (2 pts) You realize that this is a good way of making restriction enzymes, and wonder why life doesn’t make transcription factors as easy. Why they have some specificity for certain DNA sequences, yet the sequences can be degenerate? Why? Why? Please give at least two reasons.

These may be useful:

- \( \log_2 (2) = 1 \)
- \( \log_2 (3) \approx 1.585 \)
- \( \log_2 (4) = 2 \)
- \( \log_2 (5) \approx 2.322 \)
- \( \log_2 (6) \approx 2.585 \)
- \( \log_2 (7) \approx 2.807 \)
- \( \log_2 (8) = 3 \)
- \( \log_2 (a/b) = \log_2 (a) - \log_2 (b) \)
6. Chromatin interaction

6.1 (1 pt) Describe the major difference between Hi-C and ChIA-PET.

The figure below is a high resolution Hi-C heatmap for a 2.3mb region on Chr4.

6.2 (2 pts) On the 1D chromosome coordinate, draw arcs that representing long range interactions that connect interacting loci.

6.3 (2 pts) Draw how this part of the chromosome would look like in a 2D looping graph.
7. Single Cell Genomics

1. Which of the following is NOT a source of error in a single cell genomics experiment?
   A. Contamination
   B. Allelic Dropout
   C. Polymerase Reversion
   D. PCR Jackpotting

2. An important goal of new single cell library prep methods is to make nucleic acid amplification quadratic, not linear.
   A. True
   B. False

3. MALBAC has lower rates of allelic dropout than MDA
   A. True
   B. False

4. Which of the following techniques was invented first?
   A. MDA
   B. MALBAC
   C. DOP-PCR

5. In a tetraploid cell, the most common sequencing artifact due to PCR jackpotting is expected to be at what frequency?
   A. 1/2
   B. 3/5
   C. 1/4
   D. 1/8
   E. 1/16