1. Metagenomics (10 pts total)

1. (2 pts) In 2004 Venter and colleagues reported on shotgun sequencing of environmental metagenomic DNA from the Sargasso Sea. They generated over 1 gigabase of data and sampled approximately 1800 species, yet were unable to assemble the complete genome of even a single organism. The same year Banfield and colleagues reported on similar metagenomic sequencing of acid mine drainage biofilm. While Banfield generated only 76 megabases of sequence, her group was able to assemble the genomes of 5 different microbial lineages. Please explain this discrepancy between the ability of the two groups to assemble genomes from their data.

2. (1 pt) The “Great Plate Count Anomaly” refers to the finding by Staley and Konopka in 1985 that the number of bacteria from a natural environment estimated using a culture-dependent method significantly underestimates the number estimated when the sample is visualized under a microscope.

   A. TRUE
   B. FALSE

3. (1 pt) If you are attempting to create a functional metagenomic library to capture an activity encoded by a 5-enzyme gene cluster, your chance of success is likely to be higher with an average insert size of 40-50 kilobases rather than 1-3 kilobases.

   A. TRUE
   B. FALSE

4. (1 pt) A nearly complete picture of the diversity of functions encoded in a microbial community can be determined by performing a 16S rRNA phylogenetic survey of that community using one the latest generation sequencing platforms.

   A. TRUE
   B. FALSE
5. (1 pt) It is possible for two bacteria to share identical copies of the 16S rRNA gene (signifying them to be the same species) yet share less than 70% of their coding sequences.

A. TRUE
B. FALSE

6. (1 pt) Even though the number of microbial genes associated with microorganisms living in/on the human body outnumber those of the human genome almost 100-fold, the number of human cells still outnumber the number of total microbial cells associated with the human body.

A. TRUE
B. FALSE

7. (1 pt) Lateral or horizontal gene transfer is the process by which bacteria acquire new genes/functions through the acquisition and stable replication of stretches of exogenous DNA.

A. TRUE
B. FALSE

8. (1 pt) One of the advantages of a function-driven screen or selection of a metagenomic library is that its success in discovering functional genes is not dependent on prior knowledge of the sequence of the captured genes.

A. TRUE
B. FALSE

9. (1 pt) The most accurate estimates of the diversity of prokaryotic cells in a soil sample can be derived by:
   a. Making serial dilutions of the soil inoculated into rich media, plating on agar, and counting colonies over the next few days
   b. Infecting a set of live animals with the soil and a control bacterium and measuring relative animal mortality after 30 days
   c. Sequencing amplified 16S ribosomal DNA from the sample, and comparing to a database of known 16S sequences
   d. None of the above are relevant, since soil microbes are largely archaeal or eukaryotic, while prokaryotes are primarily associated with higher living organisms
2. Genetic Variations (10 pts total)

1. (1 pt) Assuming a mutation rate of $1.2 \times 10^{-8}$ per generation for single nucleotide variants (SNVs), how many new SNVs does the average human child have relative to his/her two parents? Please justify your calculations.

2. (1 pt) Approximately how many (or what fraction) of the new mutations from Question-1 are expected to be nonsynonymous single nucleotide variants that change the amino acid composition of a protein?

3. (2 pts) The mutation rate is not a constant. Give three specific examples of how mutation rates can vary in humans. For each, what is the mechanistic basis for these rate differences?

4. In class we discussed the properties of several classes of genetic variation, including single nucleotide variants, minisatellites, small indels and large CNVs.

   (i) (1 pt) Which ones of these variant types has the lowest per locus mutation rate?
(ii) (1 pt) Which one of these variant types has the highest per locus mutation rate?

(iii) (1 pt) Order these variant types by their typical genomic mutation rate (i.e. aggregating over the entire genome). For example, \( a < b < c < d \).

(iv) (2 pts) Which one of these classes of variant is the most difficult to genotype (using oligonucleotide arrays or next-generation sequencing), and why?

(v) (1 pt) On a per-variant basis, which one of these classes of variant is the most informative for personal identification and why?
3. Population Genetics (10 pts total)

You get a call from Pierre Von Nostrumhauser, a distinguished but crazy French population geneticist who has been studying an isolated group of militant hillbillies. P.V. has discovered that these anti-social yokels are genetically identical except for a single biallelic locus (DLIVRANC) that is responsible for the ability to play the banjo. Strangely, this locus is not under selective pressure. P.V., a devout banjo enthusiast, is worried that the population may lose its musical abilities, so he asks you to investigate the decay of heterozygosity. The total population size is 200 individuals, and seems to have been this size for all of time.

1. (3 pts) P.V. has sequenced the loci in 50 individuals from the Georgian population. Calculate the heterozygosity, $H$, of the population given the following initial distribution of alleles. Hint: $H_t = 1 - G_t$, where $G_t$ is the homozygosity at generation $t$.

<table>
<thead>
<tr>
<th>Number</th>
<th>Allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>D-B (banjo skills)</td>
</tr>
<tr>
<td>89</td>
<td>D-NB (no banjo skills)</td>
</tr>
</tbody>
</table>

2. (2 pts) Assuming a constant population size of 200, what is the expected heterozygosity, $H$, in 12 generations? Hint: $H_t = (1 - 1/2N)H_0$
3. (2 pts) Determine the half-life of the banjo locus, i.e. the number of generations it will take for half of the heterozygosity to decay. Remember that \( t_{1/2} = 2N^*\ln(2) \), where \( \ln(2) = 0.693 \). Discuss why P.V. should or should not be worried about the future of banjo playing.

4. (3 pts) P.V would like to know if homozygous individuals (D-B/D-B) have better banjo skills than heterozygous (D-B/D-NB) individuals. He would like to study at least 1 homozygous individual. How many individuals will P.V. have to sequence to have a 95% chance of sampling at least 1 homozygote? Remember that at Hardy-Weinberg equilibrium \( p^2 + 2pq + q^2 = 1 \)
Binomial distribution: \( P(X=n) = \left( \frac{N!}{(N-n)!n!} \right) p^n (1-p)^{N-n} \)
4. Synthetic Biology

Question 1 is worth 2 points; All other questions are worth 1 point each; ONLY 1 possible correct answer per multiple choice question:

1. Elowitz and Leibler described one of the first successful synthetic biology circuits with the design of the "Repressilator" (Nature 403, 2000). The design involved three promoters (p1, p2, p3) and three gene products which were repressors (R1, R2, R3) of those promoters. The transcriptional regulation was: p1 promotes transcription of R1, p2 promotes transcription of R2, and p3 promotes transcription of R3.
   a. Use a simple schematic to depict the setup of the Repressilator, assuming all six components mentioned are encoded on a single circular plasmid. Do not worry about any other elements on the plasmid. Label all six components and label the appropriate transcription and repression. Based on your diagram, complete the following statements:
      i. R1 represses p_
      ii. R2 represses p_
      iii. R3 represses p_

2. In a designed genetic feedback loop, promoter A drives expression of Protein B. Protein B represses promoter A. Which is the only TRUE statement below describing this system:
   a. This is a positive feedback loop
   b. This is an example of an ‘AND’ logic gate with memory
   c. Assuming no degradation of the mRNA or protein, this system will continuously oscillate between high and low expression of protein B
   d. The distribution of protein B expression levels in this system is much tighter than a system where protein B expression is unregulated

3. A hallmark of a bistable system is a continuum of possible states that can exist simultaneously but only between the system energy minimum and the system energy maximum
   a. TRUE
   b. FALSE
4. Which of the following network features were not optimized to generate an artemisin biosynthetic pathway in microbes?
   a. Ribosome binding site sequences.
   b. Enzyme active site configurations.
   c. Protein stoichiometry in enzymatic complexes.
   d. Codon compatibility with host organism.

5. Researchers have been designing synthetic gene circuits for at least a decade. Which of the following statements about gene circuit design is FALSE:
   a. The earliest synthetic circuit designs (e.g. positive and negative feedback) tried to recapitulate logic gates observed in natural systems
   b. The total number of “genetic parts” per successfully designed synthetic circuit have been steadily and rapidly increasing with the number of new circuits reported
   c. A successful synthetic negative feedback circuit yields tighter distribution of the levels of regulated reporter than an unregulated system
   d. Synthetic positive feedback can be self-perpetuating and essentially irreversible

6. You are interested in combinatorial optimization of intergenic regions in a bacterial synthetic operon, as a way of optimizing transcriptional and translational regulation of your open reading frames. Which of the following features would NOT be useful to vary in your optimization:
   a. Promoter site
   b. Ribosome binding site
   c. RNase specificity site
   d. 3' intronic splicing site

7. When a gene is introduced into a new microbial host from a genetically distinct donor, codon optimization of the transferred open-reading frame is often useful for improving gene expression in the new host. You would like to clone in a gene product which contains 14 Alanines, 6 Tyrosines, 4 Cysteines, 13 Glycines and more than 10 each of the other amino acids. Which of the following incompatibilities between the donor and the recipient is likely to be addressed by codon optimization, resulting in improved expression of your gene:
   a. The recipient lacks three of the tRNAs for Alanine that are common in the donor
   b. The recipient lacks a key tyrosine kinase that is common in the donor
   c. The recipient is deficient in disulfide formation, whereas the donor has an active periplasmic space which provides an appropriate oxidative environment for disulfide formation
   d. The recipient lacks chaperones, which have been shown to be important for protein folding in the donor
8. Name one technique for generating a diverse library of genes when you have access to a family of related genes

9. Name one technique for generating a diverse library of genes when you have access to only have one gene.
5. Massive parallel reporter assays

1. Massively Parallel Reporter Gene Assays (MPRAs) are one way to identify causal variants within eQTL. In general, MPRAs can test whether a variant is sufficient to cause an expression change but not whether it is necessary to cause a change. Why?

2. Sometimes a variant will affect gene expression in a different direction in an MPRA than the eQTL. In other words, a variant may be associated with increased expression in an eQTL, but decreased expression in an MPRA. Does this sort of discrepancy necessarily rule out the variant as being causal in an eQTL? Why?
6. Comparative Genomics (5 pts total)

The binomial distribution, \( P(X=n) = \frac{(N!}{(N-n)!})\frac{n}{n}(1-p)^{N-n} \), can be used to determine the significance of a run of \( n \) conserved nucleotides in a stretch of a multiple alignment \( N \) bases long as long as one has a good estimate of \( p \), the frequency of observing an identity in alignments of neutrally evolving DNA. (Note: \( x^0 = 0! = 1 \))

1. Your advisor, Dr. Credulous, is interested in detecting human non-coding sequences that are under positive selection. One definition of positive selection might be a sequence that has accumulated more substitutions through evolution than expected by chance. Since your advisor has not taken Bio5488 he spends three weeks scrolling through the genome browser looking for interesting sequences. One day he comes running out of his office and shows you the following 12 bp alignment:

```
1----------12
Human    ATCCGGGATCGT
Chimp    ATACGTGATAGT
      **     **   ***   **
```

Assuming that humans and chimps are 98% identical in neutrally evolving regions of the genome follow these steps to set up the formula for deciding whether you can be 95% certain that this alignment contains more substitutions than expected by chance.

(i) (1 pt) First decide what \( p \) is in this case, and what it stands for.

(ii) (2 pts) In words describe the calculation you will make. Since it will be a cumulative binomial describe the range over which you will compute, and what the result must be to be 95% confident.
(iii) (2 pts) Set up, but do not actually compute, the calculation you just described.