### SECTION 1: METAGENOMICS AND SYNTHETIC BIOLOGY (20 pts total):

1. The expansion in gene synthesis capacity (and subsequently genome-scale synthesis) has necessitated the incorporation of error-correcting techniques to achieve high sequence-fidelity. Describe three error-correcting techniques that can be used for such purposes (3 pts):

2. 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 (3 pts):
      i. R1 represses p_
      ii. R2 represses p_
      iii. R3 represses p_
   b. Their experiment began with a simple computational model of transcriptional regulation, which predicted sustained limit-cycle oscillations when four conditions were achieved. List any three (3) of the conditions (3pts):
MULTIPLE CHOICE (only 1 possible correct answer) or TRUE/FALSE (1pt each)

3. Multicellular lineages account for greater than 90% of lineages on the tree of life, as measured by rRNA.
   a. TRUE
   b. FALSE

4. The fundamental similarity between rAAV-based and CRISPR-based genome editing technologies is the introduction of a double-stranded DNA break, which then enables insertions/deletion/modifications
   a. TRUE
   b. FALSE

5. Horizontal gene transfer can only occur within the members of a specific kingdom (Bacteria, Archaea, Eukarya), and not between members of different kingdoms.
   a. TRUE
   b. FALSE

6. One of the primary advantages of a function-driven screen or selection of a metagenomic library is that its success in discovering functional genes is largely independent of prior knowledge of the sequence of the captured genes.
   a. TRUE
   b. FALSE

7. Alpha diversity measures the diversity between between two or more samples.
   a. TRUE
   b. FALSE

8. 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

9. Composition-based binning methods for DNA sequence analysis rely on comparison of gross-features (e.g. GC content) common to phylogenetic divisions (e.g. family/genus/species)
   a. TRUE
   b. FALSE
10. 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 Serines, 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 is naturally competent, while the donor primarily uses phage transduction for horizontal gene transfer.
   b. The recipient lacks four of the tRNAs for Serine that are common in the donor.
   c. The recipient is resistant to the amino-acid derivative D-cycloserine, whereas the donor is susceptible.
   d. The recipient is deficient in disulfide formation, whereas the donor has an active periplasmic space which provides an appropriate oxidative environment for disulfide formation.
   e. The recipient can utilize D-alanine as a sole carbon source, whereas the donor cannot.

11. The approximate ratio of the number of genes encoded by the human genome to the number of genes encoded by the human microbiota (all microbes in and on a human body), in one human, is:
   a. 10:1
   b. 1:100
   c. 1:1
   d. 100:1
   e. 1:10

12. Prokaryotic cells have been identified in the following human body habitats:
   a. Gut and skin only
   b. Gut only
   c. Gut and mouth only
   d. Mouth only
   e. Every habitat tested, including gut, skin, and mouth

13. In 2004, the Banfield group sequenced ~77 megabases of metagenomic DNA from acid-mine drainage, and were able to assemble 5 distinct microbial genomes from this data. In comparison, the Venter group that same year sequenced ~1000 megabases of metagenomic DNA from the Sargasso Sea, but were unable to assemble any microbial genomes from their data. The reason for this difference is:
a. Due to the limitations of working on a boat, the ability to adequately preserve microbial DNA was diminished on the Sargasso Sea research vessel, and the resultant deterioration of the sea-derived DNA contributed to poor sequence quality.
b. The Sargasso Sea data was dominated by bacteria, whereas the acid-mine data was entirely archaeal, and the lack of repetitive elements and more consistent genomic architecture in archaea makes their genomes much easier to assemble than bacteria.
c. The harsh acid-mine environment had caused a substantial reduction in species diversity and hence the ~77 megabases of sequence resulted in sufficient genome coverage of the few species present, whereas the massive diversity of organisms in the Sargasso Sea meant even ~1000 megabases of sequence was insufficient to generate sufficient genome coverage to allow assembly.
d. The low pH conditions of the acid-mine drainage caused substantial intragenomic cross-linking, allowing individual genomes to be deconvoluted from each other, whereas the Sargasso Sea lacked this feature.
e. The Venter expedition forgot to take T-Pain along.
SECTION 2: GENETIC VARIATION (10 pts total)

(a) 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.

(b) Approximately how many (or what fraction) of the new mutations from part (a) are expected to be nonsynonymous single nucleotide variants that change the amino acid composition of a protein?

(c) 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?

(d) The genomes of two "normal" human individuals differ from one another by ~4 million genetic variants, and these variants come in diverse forms. (i) Name and define at least five major classes of human genome variation. For each class of variation, (ii) estimate within an order of magnitude how many variants exist between two human individuals, and (iii) give an example of a mutational mechanism that could generate a variant of that class. (iv) Which class of genome variation has the lowest per locus mutation rate? which has the highest?
(e) You perform 50X genome sequencing on a tumor sample. This tumor is composed of multiple genetically distinct populations of cells due to the process of clonal evolution. This tumor sample is also contaminated by normal stromal cells at a level of 30%. On average, how many reads will identify a somatic variant that is present at a variant allele frequency of 0.1?

(f) Human genome structural variation can be detected using four general methods: array comparative genomic hybridization, read-depth analysis of Illumina genome sequencing data, paired-end mapping analysis of genome sequencing data (a.k.a., read-pair analysis), and split-read mapping analysis of genome sequencing data. (i) Give a 1-2 sentence description of how each method is able to detect a structural variant. (ii) Which of these four methods provides the highest genomic resolution? (iii) Which method(s) can detect balanced rearrangements such as inversions? (iv) Which method(s) can detect copy number variants (CNVs) such as deletions and duplications? (v) Which method(s) can estimate the absolute copy number of a structurally variable genomic segment?
### SECTION 3: POPULATION GENETICS

3A (1 pt) In a Wright-Fisher model without mutation or selection, heterozygosity at a locus is always expected to decrease with time

   TRUE
   FALSE

3B (1 pt) Genetic drift increases with population size

   TRUE
   FALSE

3C (1 pt) When considering just neutral variation, the expected site frequency spectrum for polymorphisms in humans is skewed toward rare events

   TRUE
   FALSE

3D (1 pt) According to Neutral Theory, the Tajima Estimator and Watterson Estimator of the population scaled mutation rate should give the same result, on average.

   TRUE
   FALSE

Suppose you have collected a general, population-based sample of 100 2-locus haplotypes for two SNP loci, A and B, that are within 5 kb of each other in an important candidate gene for disease X. Suppose the allele frequencies at locus A are \( p(A_1) = 0.2 \) and \( p(A_2) = 0.8 \), and the allele frequencies at locus B are \( p(B_1) = 0.1 \) and \( p(B_2) = 0.9 \).

3E (2 pts) What are the four expected haplotype frequencies \( p(A_1B_1), p(A_1B_2), p(A_2B_1), p(A_2B_2) \) under the assumption of independent assortment?

3F (2 pts) Now suppose instead that in your data, the \( r^2 \) between A and B is the maximum value that can possibly be attained between these two SNPs. Fill in the 2x2 table below (haplotype counts and marginal allele counts) below so that it corresponds to this maximum \( r^2 \).
3G (2 pts) What is $r^2$ for the above table? Given this value of $r^2$, would you recommend that researchers interested in testing this candidate gene for association with disease X genotype both SNPs, or is genotyping only one sufficient? If you genotype only one SNP, and it turns out that the other SNP is in fact the true disease locus, how is your power to detect disease association impacted? Hint: $r^2 = \frac{(D^2)}{(p(A_1)p(A_2)p(B_1)p(B_2))}$ where $D = P_{A_1B_1} - p(A_1)p(B_1)$. 

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>
### Section 4: Massively Parallel Reporter Assays (10 pts)

4A. (4 pts) For each of the following types of MPRA, list one advantage and one disadvantage (relative to other MPRAs) of the experimental design:

**Episomal (plasmid) library constructed from synthesized oligonucleotides**

Advantage:

Disadvantage:

**Lentiviral reporter library**

Advantage:

Disadvantage:

**STARR-seq (genomic fragments cloned in the 3’UTR of reporter gene in a plasmid)**

Advantage:

Disadvantage:

**FACS-seq/Sort-seq (Flow sorting on fluorescent reporter activity, followed by sequencing of DNA barcodes in sorted cells.)**

Advantage:
4B. (4 pts) In a study of autism spectrum disorders, your colleagues compared whole genome sequences of discordant sibling pairs, and identified 350 bi-allelic single-nucleotide polymorphisms (SNPs) that lie in a set of putative neural enhancers. They would like you to design and perform an MPRA to test which of these 350 SNPs might alter the cis-regulatory activity of the enhancers. In order to ensure that the experiment is done in a relevant cell type, your colleagues want you to perform the MPRA in mouse primary neurons, which they will gladly provide. Unfortunately, these primary neurons are not easy to transfect, so you need an assay that will work in these neurons.

Describe the design of an MPRA to test these 350 SNPs in primary mouse neurons. Be sure to include the following: 1) DNA source of your reporters, 2) Specific description of reporter library composition (total number of sequences in the library, description of negative and/or positive controls), 3) Cloning plan – type of vector and how barcodes will be used.

4C. (2 pts) You have just received the sequencing data from your MPRA of candidate autism SNPs. List two metrics that you will look at to evaluate how well the experiment worked.
The binomial distribution, \( P(X=n) = p^n(1-p)^{N-n}( N!/(N-n)!n! ) \), 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: \( x^0 = 0! = 1 \))

5A. (5 points) Your advisor is interested in detecting mammalian non-coding sequences that are under purifying selection. One definition of purifying selection might be a sequence that has accumulated fewer substitutions through evolution than expected by chance. Since your advisor has not taken Bio5488 she spends three weeks scrolling through the UCSC genome browser looking for conserved sequences. One day she shows you the following 14 bp alignment:

```
1-------------14
Human ATCCGGGATCGTAC
Mouse ATCCGTGATAGTAC
  ***** *** ****
```

Assuming that humans and mice are 80% 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 less substitutions than expected by chance.

5B (1 point) First decide what \( p \) is in this case, and what it stands for.

5C (3 points) 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.

5D (4 points) Set up, but do not actually compute, the calculation you just described.
5E (2 points) Consider the following species tree.

Suppose you are trying to identify evolutionarily conserved sequences in species A. You first align sequences from species A and B and identify blocks of sequences with high similarity. If you were trying to narrow down these conserved sequences (increase resolution) which one species should you add to the alignment to gain the most resolution? Why? Assume all of the species in the tree align well to species A.