Chemical Engineering 355
Advanced Biochemical Engineering
Spring 2018

Midterm Exam

NAME: _____

I understand and follow the Stanford Honor Code

SIGNATURE: _____

The Honor Code is an undertaking of the students, individually and collectively:

1. that they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in the preparation of reports, or in any other work that is to be used by the instructor as the basis of grading;

2. that they will do their share and take an active part in seeing to it that others as well as themselves uphold the spirit and letter of the Honor Code.

The faculty on its part manifests its confidence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid, as far as practicable, academic procedures that create temptations to violate the Honor Code.

While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to establish optimal conditions for honorable academic work.

Please do not open the exam until you are instructed to do so by the TAs. You will have 1 hour and 15 min to answer 6 questions. Several questions have a few parts; please keep track of the time and pace your work. You should work on the sections you know best first. All answers should be recorded on the exam – if you need additional space, you can use the blank paper provided by the TAs. The last page is an appendix.

You can use a calculator, and refer to any in-class or other personally written notes. A computer or laptop can be used to access the online course text book or class materials on coursework.
Not allowed: use of the internet to access any other resources

NAME: _____

1. Rates of central dogma	**14 pts**	_____
2. Homologous recombination	**12 pts**	_____
3. DNA sequencing technology	**10 pts**	_____
4. Polymerase chain reaction	**10 pts**	_____
5. Genome editing	**19 pts**	_____

TOTAL	**65 pts**	_____

1. Rates of central dogma (14 points total)

Changes to the sequence and function of the ribosome affect how all proteins in the cell are translated. This makes it extremely difficult to study potential modifications or improvements to ribosomal activity. Ribosomal evolution could be of interest for applications like the translation of unnatural amino acids or spider silk protein where the wild-type *E. coli* ribosome often tends to stall. As a result, the Jewett lab at Northwestern has developed a method to synthesize ribosomes from transcribed rRNA and purified ribosomal proteins in vitro. Modifications to ribosomal activity can be more easily studied using in vitro synthesized ribosomes because they don't affect cell growth.

Imagine you are a researcher trying to assess the activity of in vitro synthesized ribosomes. You know that *E. coli* ribosomes typically translate proteins at 16 amino acids/second, so you decide to measure the translation rate of the in vitro synthesized ribosomes to see if the translation rate is comparable.

(**A**) You first need an easy way to measure protein output. In this system, ribosomes are synthesized in the presence of mRNA for an optimized form of GFP called super-folder GFP (sfGFP). sfGFP has the same length (238 amino acids) and size (26.9 kDa) as wild-type GFP, but it folds nearly instantaneously, so it can be used as a good measure of protein translation. From the endpoint of an in vitro reaction, you find that the final sfGFP concentration of 1.00 mg/mL produces $5.00*10^5$ arbitrary units (AUs) of fluorescence in a 50.0 uL reaction. How many AUs of fluorescence are produced by a single molecule of sfGFP? **(5 points)**

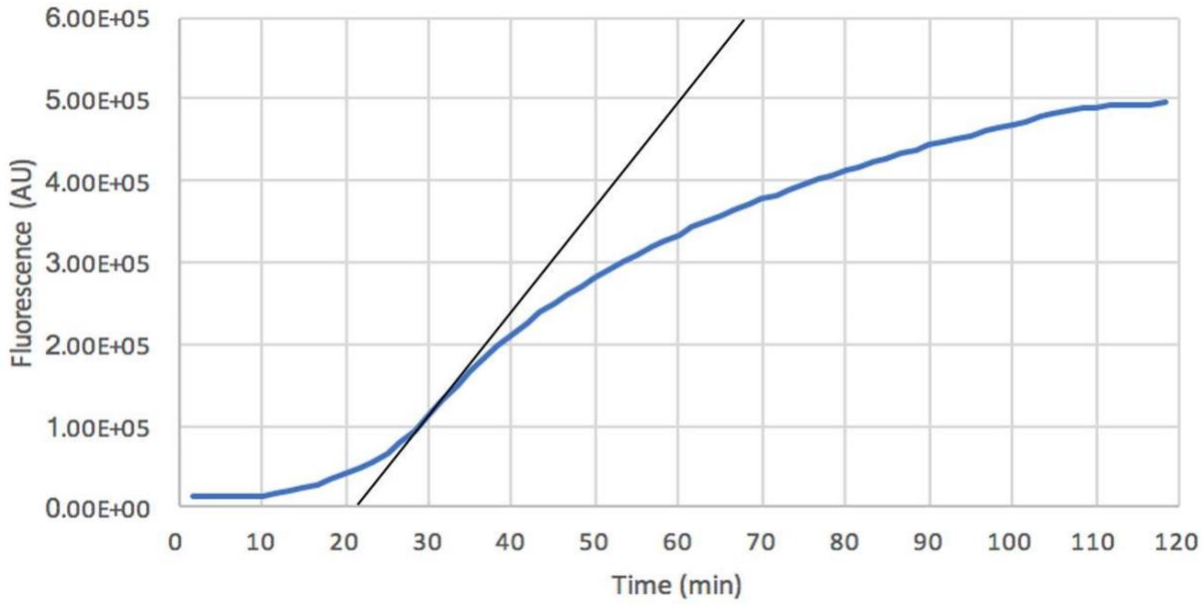
$$\frac{1\ mg\ sfGFP}{1\ mL} * \frac{50\ uL}{1} * \frac{mL}{1000\ uL} * \frac{1000\ ug}{mg} = 50\ ug\ sfGFP$$

$$\frac{50\ ug\ sfGFP}{1} * \frac{g}{10^6\ ug} * \frac{mol}{26.9 * 10^3\ g} * \frac{6.02 * 10^{23}\ molecules}{mol} = 1.12 * 10^{15}\ molecules\ GFP$$

$$\frac{500{,}000\ AU}{1.12 * 10^{15}\ molecules\ GFP} = \frac{\mathbf{4.46 * 10^{-10}\ AU}}{\mathbf{molecule\ GFP}}$$

-2 for 3 order of magnitude difference
-4 for >3 order of magnitude difference
-2 for use of average amino acid mass (not necessary for this calculation)

(**B**) In order to find the maximum rate of translation, you want to find the maximum rate of sfGFP production. The reaction has a lag phase where ribosomes are synthesized and sfGFP mRNA is produced. sfGFP production reaches a maximum for some period of time before the reaction rate decays to zero as reactants are exhausted. A graph showing how fluorescence changes with time is shown below:



From the graph, estimate the point of maximum rate of fluorescence per second and the number of sfGFP proteins produced per second at this point. **(4 points)**

Maximum rate around 30 minutes

$$\frac{600{,}000\ AU}{46\ minutes} * \frac{1\ minutes}{60\ seconds} = \mathbf{217}\ \frac{\boldsymbol{AU}}{\boldsymbol{second}}$$

Rate is an estimate, so +/- 25% on slope estimate will be accepted as long as logic is correct (range is 163 AU/second to 271 AU/second, -2 for value outside of this range). Subsequent answers based on this rate accepted as long as the logic is consistent.

$$\frac{217\ AU}{second} * \frac{molecule\ sfGFP}{4.5 * 10^{-10}\ AU} = \mathbf{4.82\ *10^{11}}\ \boldsymbol{sfGFP/second}$$

(**C**) Now that we know the sfGFP production rate, we can calculate the translation rate if we know the number of ribosomes. Once the reaction is finished, the ribosomes in the reaction are quantified, and it is found that $1.50*10^{13}$ ribosomes were present in the 50 uL reaction. Given this information, what is the translation rate of the synthesized ribosomes? **(3 points)**

$$\frac{4.82 * 10^{11}\ sfGFP}{second} * \frac{238\ amino\ acids}{sfGFP} * \frac{1}{1.5 * 10^{13}\ ribosomes} = \frac{\boldsymbol{7.6\ amino\ acids}}{\boldsymbol{second * ribosome}}$$

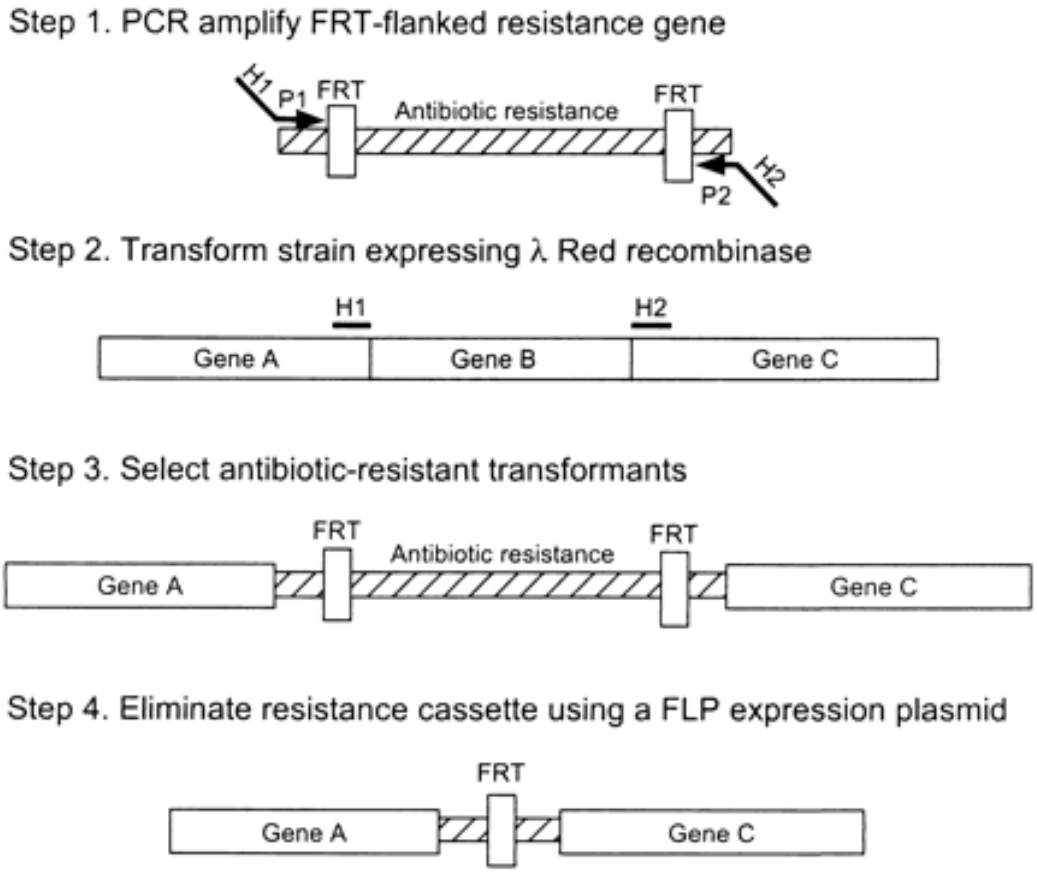
(**D**) This translation rate is lower than that of in vivo ribosomes. Give at least one reason why the translation rate would be lower than the maximum 16 amino acids per second for in vivo ribosomes. **(2 points)**

Lower activity of synthesized ribosomes compared to natural ribosomes
Inadequate aminoacyl-tRNA availability/transport
Inadequate sfGFP mRNA availability

Inadequate ATP availability
Non-functional ribosomes misfolded alongside the functional ribosomes
Some other reason that makes sense

2. Homologous recombination (12 points total)

In class we discussed the value of making a whole-genome mutant library and how it can be used to determine the function of individual genes in an organism. We also discussed how homologous recombination can be used to make these libraries, but that different strategies are required to increase the frequency of HR. One method mentioned was the expression of λ red recombinase, as described by Wanner and Datsenko in 2000 to enable the production of a mutant library in *E. coli* that contains clean knock-outs of almost every non-essential gene in the genome. As a testament to the value of such a library, this paper has been cited over 10,000 times. Please answer the following questions regarding the figure detailing the method from this paper. Note FRT/FLP operates just like the Cre/LoxP system we discussed in class, where FRT is the DNA sequence, and FLP is the recombinase.



(**A**) Our goal is to remove only gene B, leaving gene A and gene C in the genomic DNA fragment shown in Step 2. Indicate homology regions for H1 and H2 that would result in a clean deletion of only Gene B. **(3 points)**

(**B**) In the space provided in the above diagram, sketch the expected product after Step 3 **(3 points)**

(**C**) In the space provided in the above diagram, sketch the expected product after Step 4 **(3 points)**

(**D**) This method allows you to perform a single selection for double homologous recombination (in contrast to the example shown in class with a plasmid as the donor DNA). In this method, what would happen if only one recombination event took place? **(3 points)**

The genomic DNA will be split into two and the cell will not survive because E. coli has not NHEJ mechanism.

The genomic DNA splitting into to pieces and the cell dying are adequate explanations on their own.

3. DNA sequencing technologies (10 points total)

A recent paper describes the use of a new long-read sequencing technology called "Nanopore sequencing" that yields very long sequencing reads, albeit with less accuracy than existing next-generation sequencing methods. The technology works by using electrophoresis to pull a single strand of DNA through a small pore embedded in a membrane. Base-specific changes in current can be detected as the DNA is moving through the pore, allowing bases to be "read". Notably, this method sequences DNA by scanning, rather than sequencing-by-synthesis which is used by all other sequencing methods we discussed in class.

(**A**) What limits the read length for Illumina sequencing technology? Why is this not a limitation of Nanopore sequencing? **(3 points)**

Error rates get too high, and the extension and read cycles can become out of sync and bridge amplification becomes tricky for long strands (2 pts). Nanopore does not suffer from this because the read length is limited only by the translocation of the DNA strand through the pores (1 pt).

(**B**) Although Nanopore technology is relatively cheap and offers very long sequencing reads, it suffers from comparatively low accuracy. Explain what aspect of Illumina sequencing provides reads that are much more accurate than single-molecule sequencing techniques such as Nanopore or Pacbio. **(3 points)**

Clonal amplification of each DNA strand to form clusters, so as long as the error rate is DNA strand replication is rare, the majority of each clonal cluster will represent a consensus sequence that is almost perfectly accurate (2 pt). This amounts to a stronger signal-to-noise ratio (1 pt).

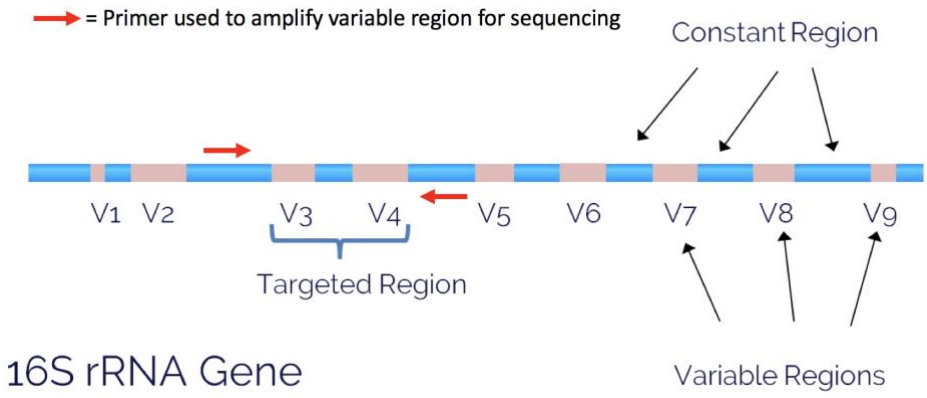
(**C**) What is the major difference between the Illumina and Roche 454 sequencing technologies in terms of base incorporation and detection. State a benefit of each technology. **(4 points)**

In Illumina, all four dNTPs are added at the same time, while in Roche 454 one dNTP is added at a time (1 pt). Detection in Illumina is the visualization of 4 different colors. Detection in Roche is presence or absence of signal (1 pt). The benefit of Illumina is that it is faster/simpler to add all 4 dNTPs at once, while the benefit of Roche 454 is that only one readout is needed (as opposed to four 4 readouts for Illumina) (2 pts).

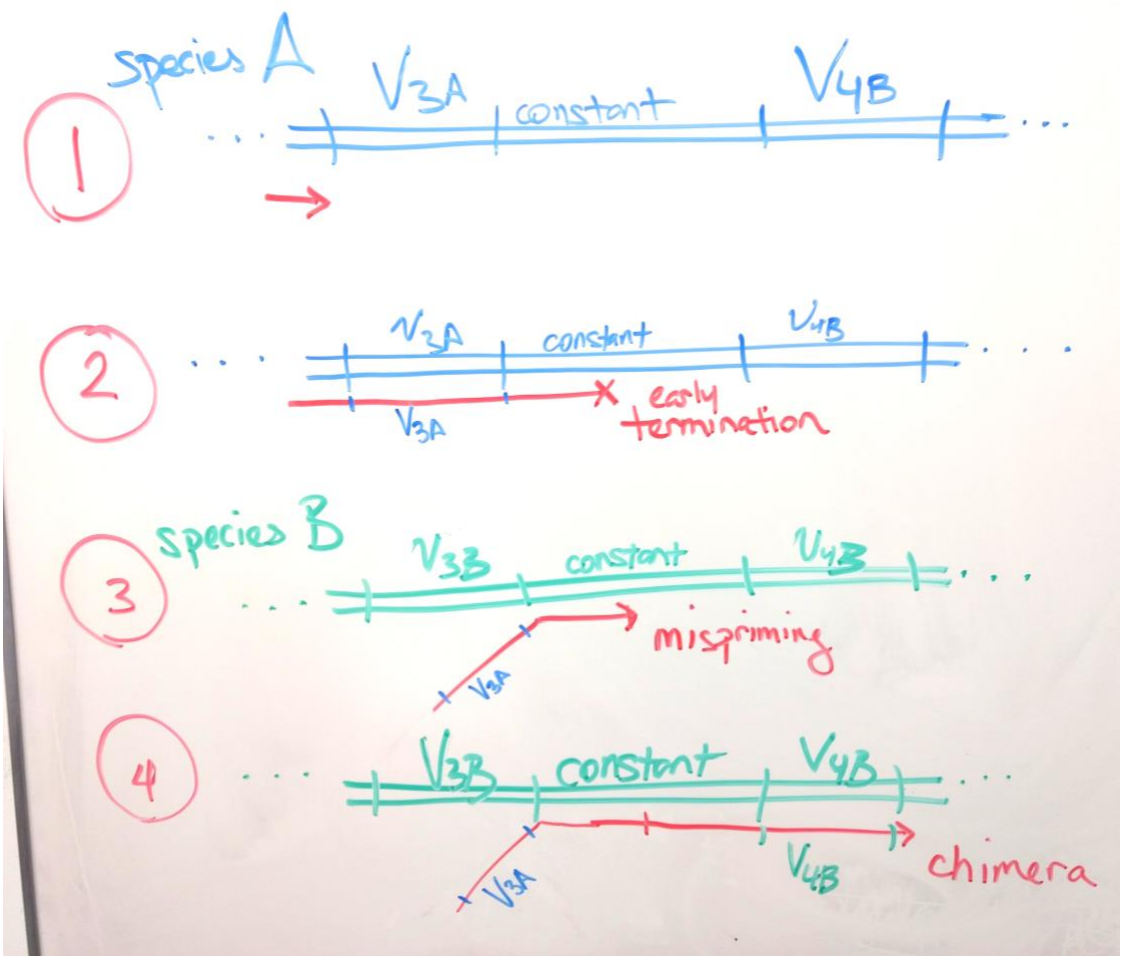
4. Polymerase chain reaction (10 points total)

Sequencing of bacterial populations has become an important way to study microbiomes; in one recent effort, researchers have used sequencing to catalog the bacterial strains that inhabit the New York City subway (they didn't find anything too surprising– mostly human-associated commensal organisms). In order to determine which bacterial genera are present in a sample of environmental DNA, researchers typically amplify and sequence the genes in the population that encode the 16S ribosomal RNA, which is part of the 30S subunit of the ribosome. This gene is ubiquitous in bacteria and has both highly conserved regions that make it easy to amplify, and a variable region that is typically unique to each group of closely related bacteria (see picture, below). Termed "16S" sequencing, this method is commonly used to determine the identities of bacteria in a community and their relative abundance.

Early in the adaptation of this method, researchers began to notice a strange phenomenon. Sequencing of a defined 'test' bacterial community that consisted of 20 known, characterized bacterial strains yielded sequencing reads for >100 putative 16S genes that were not part of the original population. Once contamination was ruled out, it was discovered that the "new" species were actually chimeric 16S genes composed of part of the 16S genes of species that were known to be in the sample. In fact, most data analysis pipelines now include an algorithm called "chimeraslayer" that removes these sequences.



(**A**) Provide one possible mechanism for how a chimeric sequence could have appeared in the sequencing of amplified 16S genes from an environmental DNA sample. Note that polymerases can sometimes fall off a template before a complete copy is made. Please use a drawing to support your answer. **(6 points)**



Replication starts on species A, then falls off in a constant region (2 pts). The partially-replicated DNA strand then misprimes onto species B in the same constant region (2 pts). Replication continues on species B, giving an A/B chimera (in the correct order; i.e., V3A-constant-V4B, not V3A-constant-V3B) (2 pts).

(**B**) Let's imagine a simple case where you have an artificial community comprised of 20 known strains of bacteria, all in equal abundance. After 30 rounds of PCR and DNA sequencing, how could you easily filter out chimeric reads based only on read quantification? **(4 points)**

Because chimeras rarely occur and we know the bacterial strains are in equal proportions, you can assume the least abundant read groups are chimeras and discard those (4 pts).

5. Genome editing (20 points total)

The engineered CRISPR-Cas9 system discussed in class can be used to cut any 23 bp DNA site of the following form:

5'-NNNNNNNNNNNNNNNNNNNNGG-3' or **5'-$(N)_{20}$NGG-3'**
3'-NNNNNNNNNNNNNNNNNNNNCC-5' equivalently: **3'-$(N)_{20}$NCC-5'**

where N can be any nucleotide and the terminal NGG represents the PAM.

(**A**) Suppose we choose some particular sequence to target (that is, we set specific nucleotides in place of all the Ns above) using an engineered Cas9 system. How frequently would we expect to cut a completely random DNA sequence, assuming 25% each of G, C, A and T? **(4 points)**

We have a 23 bp non-palindromic recognition site, with all bases equally probable. Hence, the cutting frequency is $2*(0.25)^{22}$ = **$1.137x10^{-13}$**. The factor of 2 accounts for the fact that the site is non-palindromic, so that the complementary sequence

5′–CCNNNNNNNNNNNNNNNNNNNN–3′
3′–GGNNNNNNNNNNNNNNNNNNNN–5′

is also recognized.

Why is it 22 instead of 23? 20 bases need to match with the sgRNA. The GG in the DNA must be there to be recognized by the Cas9 protein. The N in the PAM site is not specific! The N from the PAM site can be anything and Cas9 can still cut. (-1 point for no complementary strand; -1 point for wrong exponent; +1 point for showing correct thinking)

(**B**) We want to use the CRISPR-Cas9 system to help develop a cure for Huntington's disease, which is caused by a faulty copy of the huntingtin gene, located on human chromosome 4. While the exact role of the Huntingtin protein is unknown, sequencing has revealed that its coding DNA sequence contains a poly-CAG codon repeat region, corresponding to a poly-glutamine (Q = glutamine) repeat in the mature protein sequence:

Coding strand: **5' - …AAG TCC TTC CAG CAG CAG … CAG CCG CCA … - 3'**

X CAG repeats

Translation: **N - … K S F Q Q Q … Q P P … - C**

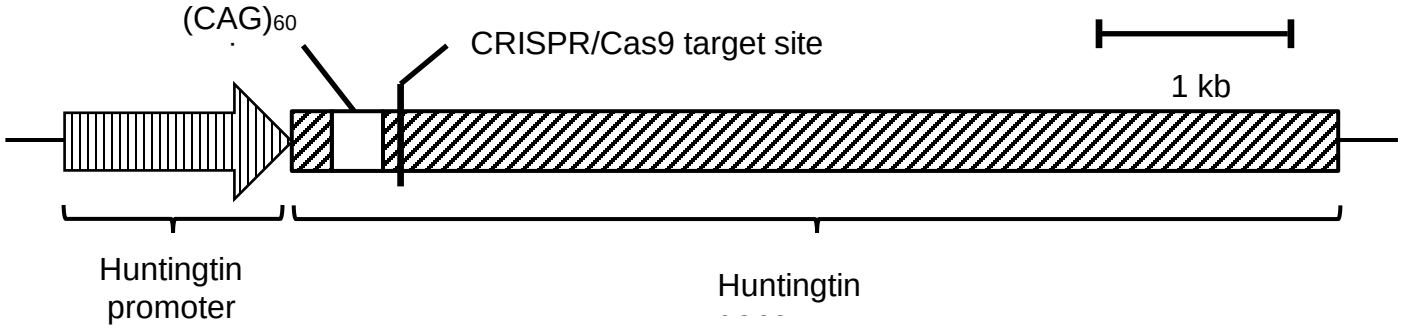
X Q (glutamine) residues

Different individuals have different numbers (X) of CAG/Q repeats. The number of repeats determines whether or not they will suffer from Huntington's disease:

# of repeats (X)	**Disease outcome**

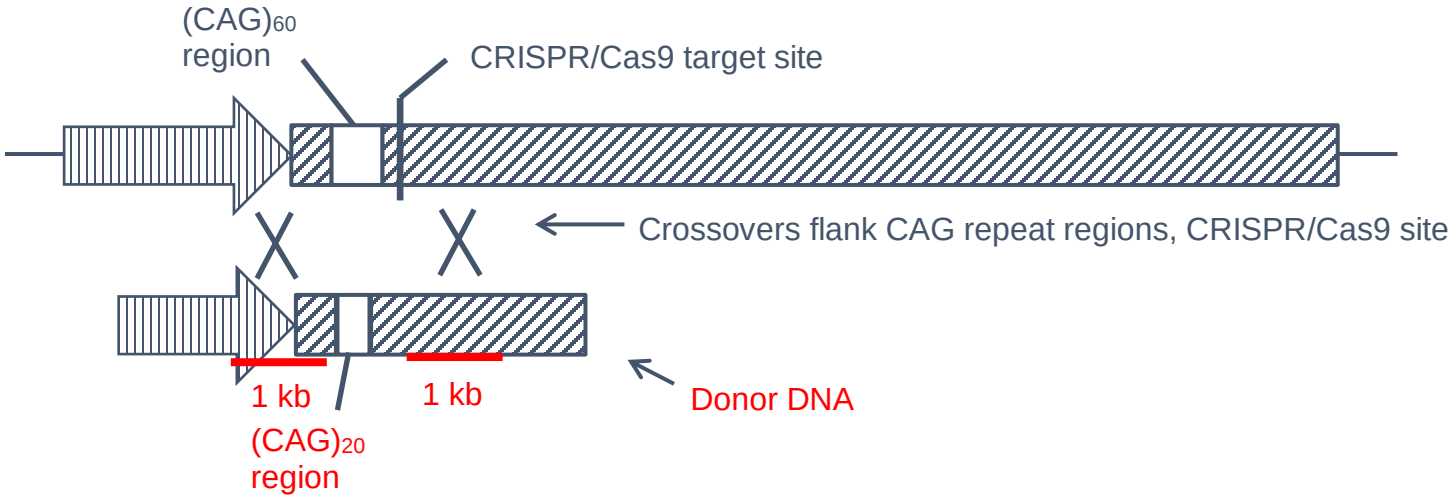
20-35	Will not be affected
36-39	May be affected
>39	Will be affected

Our idea is to use CRISPR/Cas9-targeted genome editing to replace a disease-causing huntingtin copy containing 60 CAG repeats with a normal copy, containing only 20 CAG repeats. Below is a schematic of the genomic region we are targeting, along with a scale bar. There is a suitable CRISPR/Cas9 site just downstream of the CAG repeat region, where we can cut to promote double crossover:



Sketch the linear donor DNA we would introduce to repair this gene using homologous recombination and indicate approximately where the crossovers would need to occur to get repaired. Assume we need at least ~1 kb on either side of the cut site to get efficient recombination; no need to sketch any of the CRISPR/Cas9 machinery – assume this is already in place. **(6 points)**

Sketch:



- -2 pts for not labeling crossover regions (-1 pt for not indicating crossover itself)
- -2 pts for not indicating length of HR regions (-1 pt for only showing promoter)
- -3 pts for showing crossovers only as defined points rather than regions
- -2 pts for not showing switching of CAG regions
- Various amounts of partial credit given depending on how off your construct was (e.g. whether the donor had a $(CAG)_{20}$ region, whether there was extraneous sequence, or whether it was clear that the flanking regions were at least 1 kb long)

(**C**) When the genome is repaired with the donor fragment, we want to ensure that the repaired sequence can no longer be recognized by the CRISPR sgRNA. What modification should we make in the donor DNA to prevent any subsequent double strand breaks in the repaired gene? **(3 points)**

We should prevent crossovers between donor DNA and genomic DNA, so we should change the CRISPR recognition site to avoid making a double strand break again. We can do so by using synonymous codons for the part of the gene that the CRISPR site is in.

-1 point for not specifying that it has to be a synonymous codon change

(**D**) We employ the above strategy in a cell line carrying the faulty huntingtin gene, and sequence the vicinity of the CAG repeat region in 10,000 cells that have undergone our treatment. Roughly 90% still contain 60 CAG repeats, while ~9% are "cured" and contain only 20 CAG repeats. Curiously, around 1% of the cells we sequence contain an intermediate number of CAG repeats – more than 20, but less than 60. What happened in this 1% of cases? Where might have one of the crossovers occurred? **(3 points)**

The crossover likely occurred between the two CAG repeat regions, which are homologous, leading to an intermediate number of CAG repeats.

(**E**) How could we modify our donor DNA to prevent undesired cases like those described in part D? Note that we still want our repaired protein to contain 20 consecutive glutamine residues. (*Hint*: Some of the information in the appendix might be useful here) **(3 points)**

Replace the $(CAG)_{20}$ region on the donor DNA with a $(CAA)_{20}$ region, which still codes for glutamine. This will eliminate the sequence homology between the repeat regions on the donor and target, effectively preventing crossovers from occurring here.

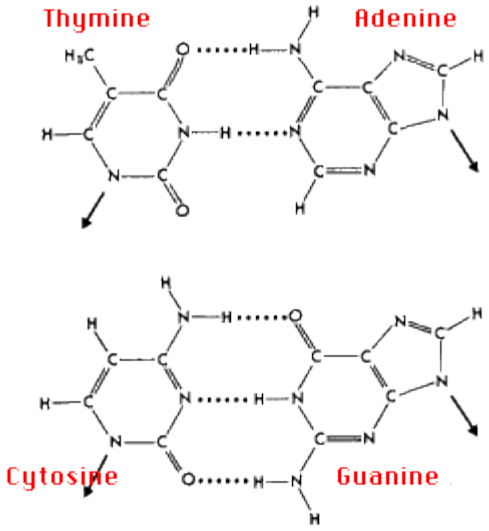
Partial credit was given for answers that would make the intermediate cases less frequent but not prevent them, such as extending the length of the flanking regions beyond 1kb.

Appendix

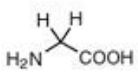
The standard genetic code:

	U	C	A	G	
U	UUU Phe - F UUC UUA Leu - L UUG	UCU Ser - S UCC UCA UCG	UAU Tyr - Y UAC UAA stop UAG stop	UGU Cys - C UGC UGA stop UGG Trp - W	U C A G
C	CUU Leu - L CUC CUA CUG	CCU Pro - P CCC CCA CCG	CAU His - H CAC CAA Gln - Q CAG	CGU Arg - R CGC CGA CGG	U C A G
A	AUU Ile - I AUC AUA AUG Met - M start	ACU Thr - T ACC ACA ACG	AAU Asn - N AAC AAA Lys - K AAG	AGU Ser - S AGC AGA Arg - R AGG	U C A G
G	GUU Val - V GUC GUA GUG	GCU Ala - A GCC GCA GCG	GAU Asp - D GAC GAA Glu - E GAG	GGU Gly - G GGC GGA GGG	U C A G

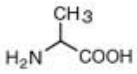
DNA base pairing:



Small

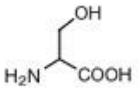


Glycine (Gly, G)
MW: 57.05

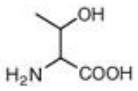


Alanine (Ala, A)
MW: 71.09

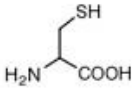
Nucleophilic



Serine (Ser, S)
MW: 87.08, pK_a ~ 16

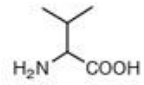


Threonine (Thr, T)
MW: 101.11, pK_a ~ 16

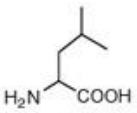


Cysteine (Cys, C)
MW: 103.15, pK_a = 8.35

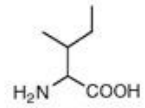
Hydrophobic



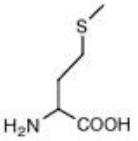
Valine (Val, V)
MW: 99.14



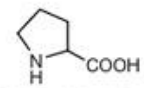
Leucine (Leu, L)
MW: 113.16



Isoleucine (Ile, I)
MW: 113.16

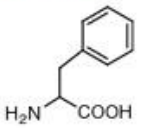


Methionine (Met, M)
MW: 131.19

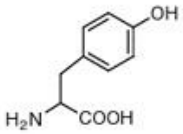


Proline (Pro, P)
MW: 97.12

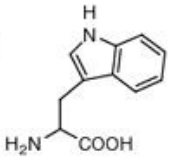
Aromatic



Phenylalanine (Phe, F)
MW: 147.18

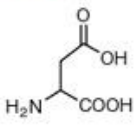


Tyrosine (Tyr, Y)
MW: 163.18

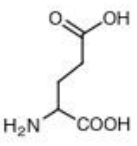


Tryptophan (Trp, W)
MW: 186.21

Acidic

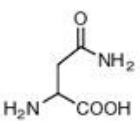


Aspartic Acid (Asp, D)
MW: 115.09, pK_a = 3.9

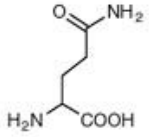


Glutamic Acid (Glu, E)
MW: 129.12, pK_a = 4.07

Amide

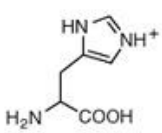


Asparagine (Asn, N)
MW: 114.11

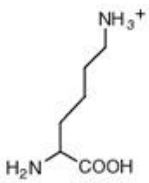


Glutamine (Gln, Q)
MW: 128.14

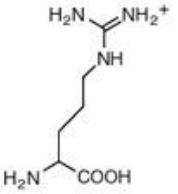
Basic



Histidine (His, H)
MW: 137.14, pK_a = 6.04



Lysine (Lys, K)
MW: 128.17, pK_a = 10.79



Arginine (Arg, R)
MW: 156.19, pK_a = 12.48