CHEMENG 355 Spring 2018

Problem Set 2 Due April 26, 2018

Problem 1: Error-Prone PCR

When performing PCR, it is important to ensure fidelity in DNA replication to obtain accurate products. Let's investigate how mutation rates over the course of replication impact PCR accuracy. Using a high-fidelity DNA polymerase, we plan to amplify a 1000-bp gene from plasmid DNA, starting with 10 pg (ca. 10 million copies) of double-stranded template DNA. For simplicity, you may assume that the polymerase always terminates replication at the end of gene of interest.

a) Assume the polymerase makes no mistakes. Derive a formula for the total number of PCR amplicons that are produced from *n* copies of template dsDNA after *k* PCR cycles (i.e., k = 1 after 1 PCR cycle, where one cycle encompasses denaturation, annealing, and extension).

Suppose a distracted researcher uses manganese chloride ($MnCl_2$) instead of magnesium chloride ($MgCl_2$). The Mn^{2+} causes the polymerase to make mistakes much more frequently. Unintentionally, this researcher thus performs error-prone PCR, which is commonly used to construct mutant libraries.

b) For a single cycle of PCR, what is the probability that a completely correct copy of the gene is generated?

c) Assume the polymerase makes a mistake once every 10000 nucleotide additions ($\mu = 10^{-4}$) in the presence of Mn^{2+}. Derive a formula for the total number of correct PCR amplicons that are produced from *n* copies of template dsDNA after *k* PCR cycles. *Hint:* assume that each nucleotide addition is a statistically independent event.

d) Using your result from part b, write the formula for the total number of mutated PCR amplicons produced from *n* copies of template dsDNA after *k* PCR cycles, where a mutated strand is defined as having one or more errors in the DNA sequence.

e) What is the percentage of mutated strands of DNA after *k* = 1 cycle? 5 cycles? 35 cycles?

f) How does this percentage of mutated strands after 35 cycles compare to a typical PCR reaction involving $MgCl_2$ and a high-fidelity polymerase ($\mu = 4.4 \times 10^{-7}$)?

g) (Bonus) Using $\mu = 10^{-4}$, after 20 cycles of PCR, on average, how many strands will contain exactly 3 errors? Approximate your answer to 3 significant figures.

Solutions (point allocations shown below explanations):

A) The amount of DNA doubles in each cycle, but we need to remove the original template strands:

$$\boldsymbol{n_o(2^k - 1)}\ (1\ point)$$

(-0.5 points for not subtracting out template)

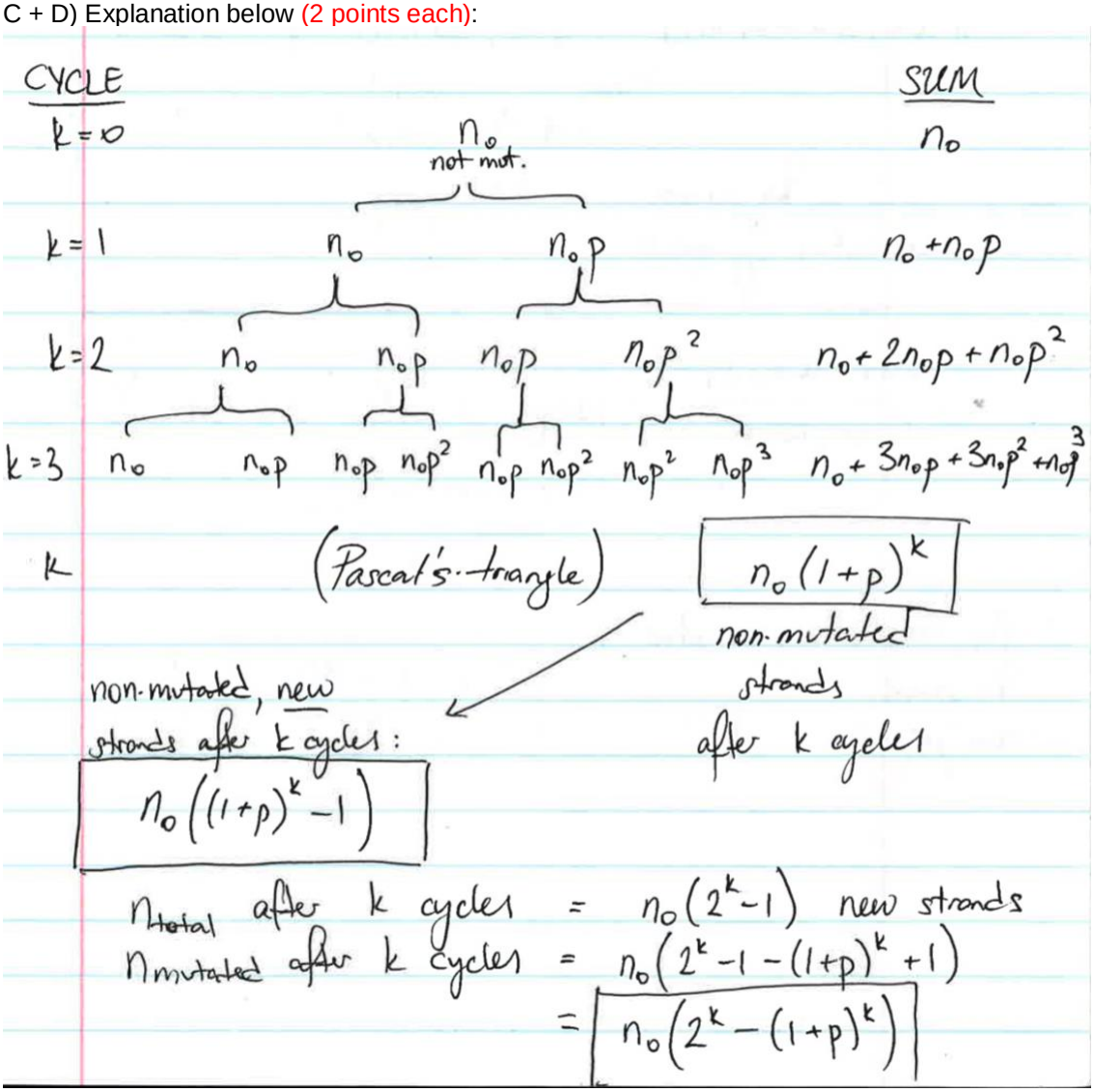
Note the simplifying assumption of the polymerase terminating replication at the end of the gene allows us to remove the -k term that would otherwise be present.

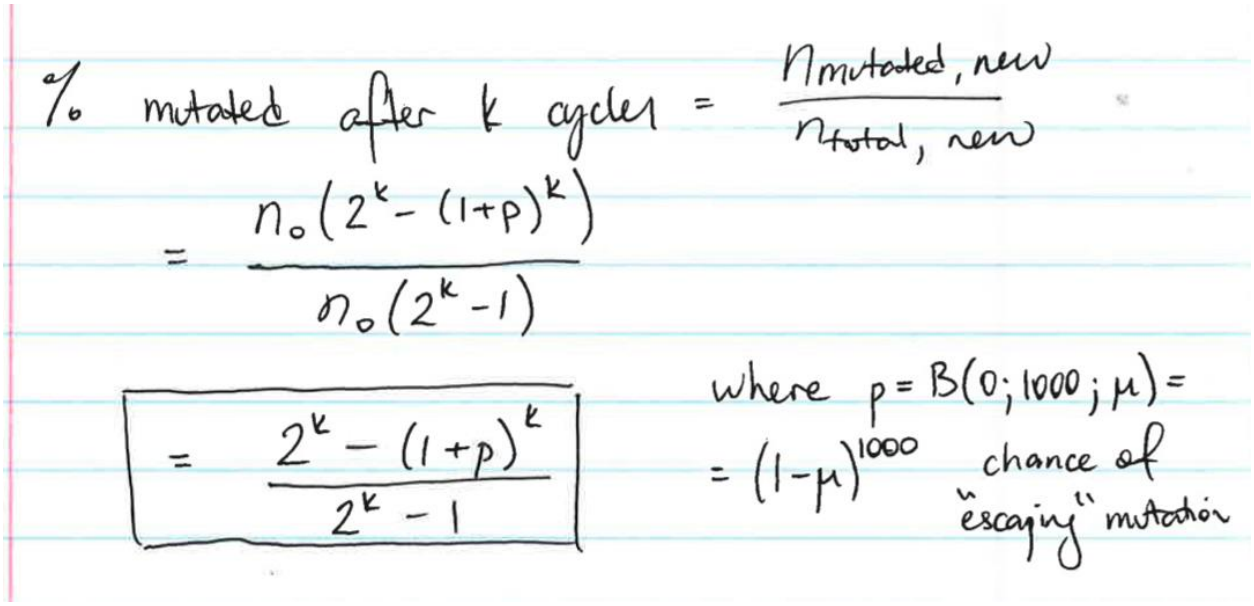
B) Each base pair insertion is independent.

$$P(1\ bp;\ incorrect) = 10^{-4}$$
$$P(1\ bp; correct) = 1 - 10^{-4} = 0.9999$$

$$P(1000\ bp; correct) = (0.9999)^{1000} = \mathbf{0.9048}\ (1\ point)$$

C + D) Explanation below (2 points each):





E) plug the cycle number k into the formula derived above:

$$Percentage\ Mutated = \frac{2^k - (1+P)^k}{2^k - 1}$$

Percentage mutated after:

- **1 cycle = 10%** (1 point)
- **5 cycles = 22%** (1 point)
- **35 cycles = 82%** (1 point)

Common mistake to drop the – 1 in the bottom of the formula. This doesn't affect the answer for cycle 35 but does for cycles 1 and 5.

F) (1 point) Comparison to high-fidelity polymerase: 82% versus **0.767%** or over 100-fold.
Use the same formula as in part E, but p changes with the new polymerase!

G) (up to 3 points)

Problem 16

- Note: Many ways to get to 3x mutated strand. Can mutate away from 3x mutated

Easiest to just track it in Excel. First, must ask, if I copy a strand, what is the probability that I add 0, 1, 2, or 3 mutations.

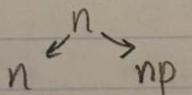
$P_0 = (1-\mu)^{1000} = 0.90483$

$P_1 = (1-\mu)^{999}(\mu)\binom{1000}{1} = 0.09049$

$P_2 = (1-\mu)^{998}\mu^2\binom{1000}{2} = 0.00452$

$P_3 = (1-\mu)^{997}\mu^3\binom{1000}{3} = 0.00015$

Now, let's see what happens to the population after 1 cycle:



Percentage of strands unmutated: $\frac{n+np_0}{2n} = 0.9524$

Percentage of strands mutated: $\frac{np_1}{2n}$ or $\frac{np_2}{2n}$ or $\frac{np_3}{2n}$

Note: In previous parts of this problem, we omit the template and focus on amplicons. We won't do that here. Why? We are using a recursive function where the end of each cycle is template for the next cycle. So template is important, but the actual template will be irrelevant 20 cycles in.

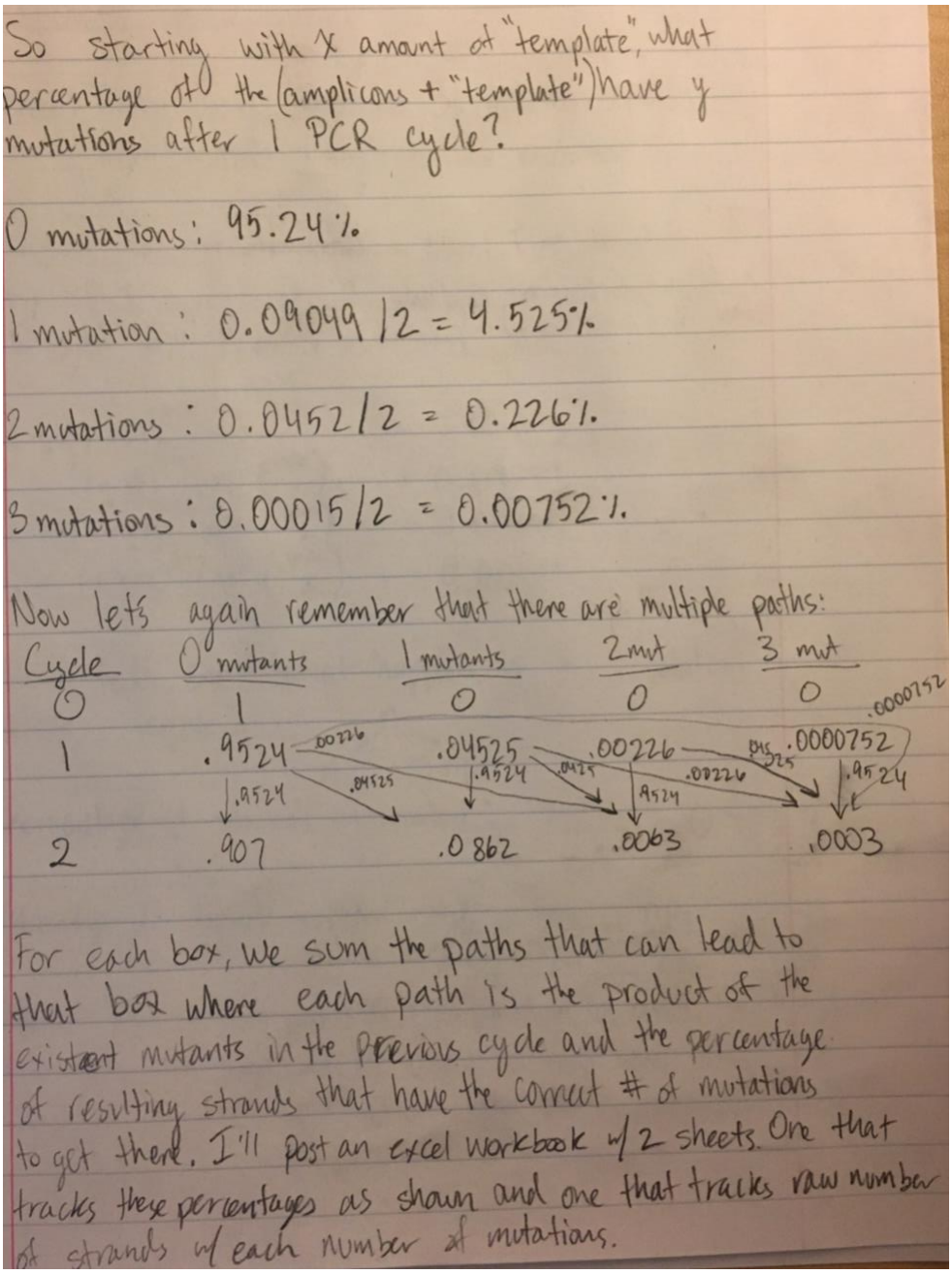


Table with proportions:

Round	Unmutated Count	Once mutated Count	Twice mutated Count	Thrice Mutated Count
0	1	0	0	0
1	0.952416447	0.045246169	0.002260272	7.51992E-05
2	0.907097088	0.086186392	0.006352657	0.000347779
3	0.863934186	0.123128005	0.012000265	0.000881682
4	0.822825127	0.15635885	0.018953047	0.001725964

5	0.783672184	0.186148425	0.026985641	0.002916679
6	0.746382277	0.212748986	0.035895384	0.004478568
7	0.710866756	0.236396572	0.045500458	0.006426588
8	0.67704119	0.257311981	0.055638176	0.008767287
9	0.644825165	0.275701683	0.066163393	0.011500031
10	0.614142092	0.291758686	0.076947029	0.01461811
11	0.584919029	0.305663348	0.087874707	0.018109723
12	0.557086503	0.317584145	0.098845488	0.02195886
13	0.530578348	0.327678393	0.109770702	0.026146078
14	0.505331545	0.336092929	0.120572865	0.0306492
15	0.481286075	0.34296475	0.131184684	0.035443925
16	0.458384773	0.348421619	0.14154813	0.040504367
17	0.436573197	0.352582636	0.151613585	0.045803534
18	0.415799493	0.355558766	0.161339059	0.051313736
19	0.396014275	0.357453351	0.170689466	0.057006948
20	0.377170509	0.358362579	0.179635949	0.062855121

Table with number of strands - basis is to start with 10 million, accepted although to start with 20 million is more accurate given 10 million dsDNA (below):

Cycle	0 mutations	1 mutation	2 mutations	3 mutations
0	10000000			
1	19048300	904900	45200	1500
2	36283773	3447361	254081.03	13894.786
3	69114420	9849956	959935.56	70483.637
4	131651220	25016655	3032233.7	276012.88
5	250773194	59565595	8634710.5	932967.41
6	477680303	1.36E+08	22971241	2865351.7
7	909899772	3.03E+08	58236072	8223747.3
8	1.733E+09	6.59E+08	142422759	22438756
9	3.301E+09	1.41E+09	338730507	58867131
10	6.289E+09	2.99E+09	787876791	149658976
11	1.198E+10	6.26E+09	1.8E+09	370816645
12	2.282E+10	1.3E+10	4.048E+09	899273352
13	4.346E+10	2.68E+10	8.992E+09	2.142E+09
14	8.279E+10	5.51E+10	1.975E+10	5.021E+09
15	1.577E+11	1.12E+11	4.298E+10	1.161E+10
16	3.004E+11	2.28E+11	9.276E+10	2.654E+10
17	5.722E+11	4.62E+11	1.987E+11	6.003E+10
18	1.09E+12	9.32E+11	4.229E+11	1.345E+11
19	2.076E+12	1.87E+12	8.948E+11	2.988E+11
20	3.955E+12	3.76E+12	1.883E+12	6.59E+11

20 million strand basis:

Cycle	0 mutations	1 mutation	2 mutations	3 mutations
0	20000000			
1	38096600	1809800	90400	3000
2	72567546	6894722	508162.06	27789.572
3	138228840	19699912	1919871.12	140967.274
4	263302440	50033310	6064467.4	552025.76
5	501546388	119131190	17269421	1865934.82
6	955360606	272000000	45942482	5730703.4
7	1819799544	606000000	116472144	16447494.6
8	3466000000	1318000000	284845518	44877512
9	6602000000	2820000000	677461014	117734262
10	1.2578E+10	5980000000	1575753582	299317952
11	2.396E+10	1.252E+10	3600000000	741633290
12	4.564E+10	2.6E+10	8096000000	1798546704
13	8.692E+10	5.36E+10	1.7984E+10	4284000000
14	1.6558E+11	1.102E+11	3.95E+10	1.0042E+10
15	3.154E+11	2.24E+11	8.596E+10	2.322E+10
16	6.008E+11	4.56E+11	1.8552E+11	5.308E+10
17	1.1444E+12	9.24E+11	3.974E+11	1.2006E+11
18	2.18E+12	1.864E+12	8.458E+11	2.69E+11
19	4.152E+12	3.74E+12	1.7896E+12	5.976E+11
20	7.91E+12	7.52E+12	3.766E+12	1.318E+12

Problem 2: Searching Genomic Libraries [10 points total]

Tardigrades, also known as "water bears" are some of the toughest creatures you may have never heard of. These small aquatic species have an extraordinary ability to withstand extreme stresses, which include extreme temperatures (~-273 °C to ~100 °C), pressures, complete immersion in organic solvents, irradiation, and even being expelled into open space by the impact of a comet. (Hashimoto *et al*., Nat. Comm., 2016, 7, 1-14)

You are a grad student who has had the good fortune to have discovered a new, highly stress-resistant tardigrade species from a sample from a pond outside of your lab. It is known that much of the tardigrade's stress resistance derives from a protective protein coating on the surface of the tardigrade body.

Obviously, being able to determine the gene encoding for this rare protective protein would be of interest to many human scientists. You decide to create a gene library using the tardigrade genomes that you isolate from your pond sample.

a) [2 points] The size of the tardigrade genome is 100 Mbp. Let's say you have materials to generate 10,000 clones. How big should your genome fragments be to have a 99% probability that the protein coat-coding gene is contained in this library?

b) [4 points] One tacit assumption here is that the fragment length is significantly longer than the length of our target gene. If the fragments are shorter, the chance of a fragment containing only a portion of the gene is higher. If we do not make this assumption, and let G be the length of the target gene (and *g* be the ratio of the target gene length to length of the genome), how does the library equation change?

c) [4 points] Coverage (C) is defined as the average number of times a given nucleotide from the genomic sequence is represented in a genomic library and is therefore simply C = N*F/L, where N is the number of clones, L = the length of the genome, and F is the average length of each fragment. For 10x coverage of the tardigrade genome, what is the probability that a nucleotide is represented in the fragment pool?

   Hint 1: Answer is independent of genome length
   Hint 2: $\ln(1 + x) \approx x$ for $|x| << 1$

d) [+2 points] Bonus: These rare proteins are certainly useful, but it might not be immediately obvious how one would construct a functional assay to know that they had been produced in a clone. Think of two possible ways that one could do this based on information given in the problem statement.

Solutions:

a)

$$N = \frac{\ln(1-P)}{\ln(1-f)}$$

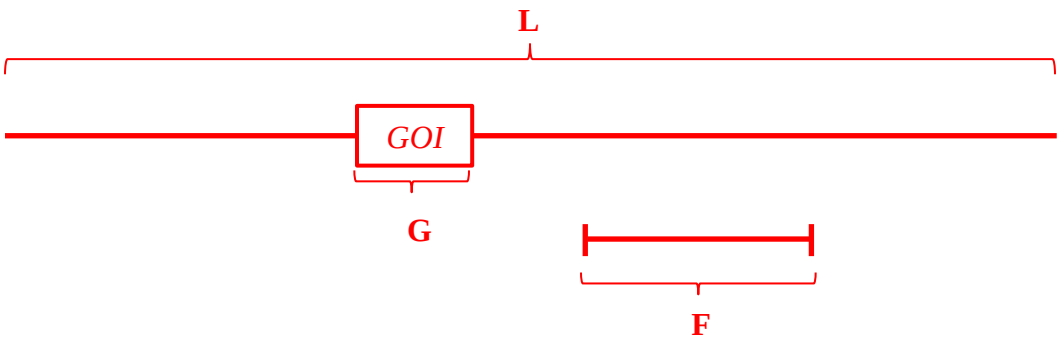
$$f = 1 - (1-P)^{\frac{1}{N}}$$

$$f = 1 - (1-0.99)^{\frac{1}{10^4}}$$

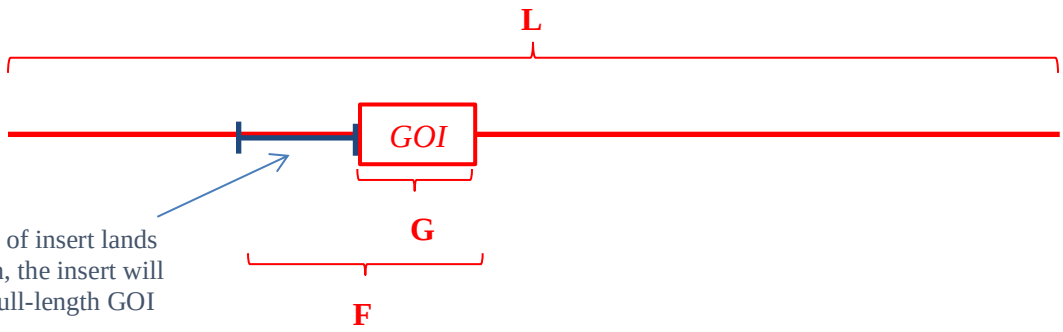
$$f = 4.6\text{x}10^{-4}$$

$$\boxed{F = 4.6\text{x}10^{4}\ \text{bp}}$$

b) We'll need to think about how the above equation is derived. Suppose we have a genome (length = L), a gene of interest (length = G) and some average insert size (length = F), sketched below:



Conceptually, the library generation procedure can be described as follows: At random, we select a subsequence of size F from the genome, and repeat N times. The subsequence will contain our full-length GOI only if the left border of this subsequence is located anywhere between the left border of the GOI or at most F – G bases to the left of this border – otherwise, the GOI will be either truncated, or not contained in the insert at all:



If left border of insert lands in this region, the insert will contain the full-length GOI

The probability that any randomly selected insert contains full-length GOI is therefore (F – G)/L. Now, the probability that *at least one* of N such inserts contains the full-length GOI can be calculated using some basic probabilistic reasoning:

P(at least one of N inserts contains GOI) = 1 – P(none of N inserts contains GOI)

But

P(none of N inserts contains GOI) = [P(random insert does not contain GOI)]^N

=[1 – P(random insert contains GOI)]^N

=[1 – (F – G)/L]^N,

from above. Therefore,

P(at least one of N inserts contains GOI) = p = 1 – [1 – (F – G)/L]^N

Realizing that F/L = *f* and G/L = *g*, after rearranging, we obtain the following:

$$\boldsymbol{N = \frac{\ln(1-p)}{\ln(1-f+g)}}$$

Answers that were not completely simplified were not considered completely correct. The F, G and L are all substantially larger than 1, so F + 1 = F; this was accepted. Also, G is still substantially smaller than L, so L + G = L. These assumptions are tacitly made when formulating the library equation and changing the assumption that F and G are comparable in size should NOT affect those assumptions.

c)

Let L = 1

$$N = \frac{\ln(1-P)}{\ln(1-f)} \approx \frac{\ln(1-P)}{-f}$$

$$Nf \approx -\ln(1-P)$$

$$C = 10 \approx -\ln(1-P)$$

$$\boldsymbol{P = 1 - \frac{1}{e^{10}} = 0.99995}$$

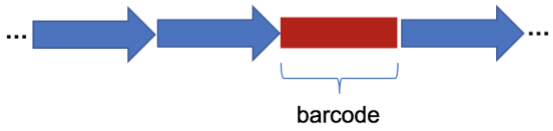
d) Use of organic solvent washing, temperature-based degradation, etc to isolate protein of interest. Variety of answers focusing on assaying the protein itself accepted. It is unlikely that the protein from a foreign organism would function to coat entire bacterial cells, so answers that only focused on bacterial host survival did not receive bonus credit.

Problem 3: Preparing DNA Libraries by PCR [10 points total]

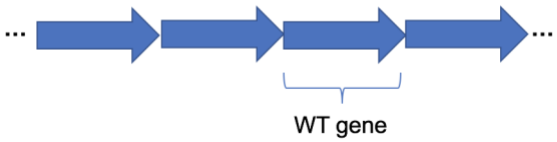
In class we discussed how a genomic library could help you find a specific gene of interest (e.g. the gene that encodes for GFP). Another way to find genes that have a certain function is to use

a mutant library. One example of a mutant library is the yeast deletion collection (produced by the Stanford Genome Technology Center) which consists of 5000 unique yeast strains that are identical except that each contains a different single ~~null~~ ~~mutation~~ (deletion). In this collection, there is at least one strain for almost every non-essential gene in the yeast genome where the wild-type gene has been individually deleted and replaced with a unique known barcode (a synthetic DNA sequence). Barcodes can be read by PCR and sequencing to determine which gene has been deleted in each particular strain.

DNA from an example mutant strain, where a single gene has been replaced with a synthetic DNA barcode sequence (in red):
(we'll discuss how genes can be edited e.g. to make a mutation and/or to add a barcode sequence in future lectures)



DNA from Wild Type (WT) yeast strain (not mutated):



Researchers have used this collection for over a decade to study the effects of perturbing individual genes on yeast physiology. Suppose you are interested in determining which genes are important for the growth of yeast at higher temperatures as an example of a stress condition.

In the yeast genome, there are genes that are crucial for survival at high temperature. Cells in which these genes are deleted will not grow well under this stress condition compared to normal growth conditions. If we use PCR and sequencing to determine the barcodes present in the strains that grow in the normal conditions compared to those that grow in the stress condition, it should allow us to find genes that are important for growth at high temperatures.

a) [3 points] You begin by taking an aliquot of the yeast deletion collection. You need to ensure that you gather enough cells such that all library members are well-represented. What is the minimum number of cells you need if 99% of all null mutant strains is each to appear at least once in the final library pool?
(Hint: to do this, write a few lines of computer code or use a software tool to simulate randomly picking from a list of 5000 library members N times, where N is an initial guess at the number of cells you need, and adjust your guess for N to get close to 99%.)

b) [2 points] We are interested in quantifying the abundance of each library member (i.e. each barcode) in the cell population before and after exposure to the stress condition. To amplify the barcodes in each pool, we would like to begin with, on average, 1000 copies of each barcode to ensure that inefficiency in PCR and sequencing does not alter each member's representation. What concentration of total library DNA is needed to attain those 1000 copies? Assume these copies are contained in a volume of 50 microliters.

c) [2 points] Over-amplifying a DNA library can cause depletion of primers, which leads to the formation of chimeric products and other undesirable effects. Practically, this means limiting the final DNA concentration achieved by PCR to less than 1 nM, which is enough for subsequent sequencing. How many cycles of PCR are necessary to amplify the initial DNA to this concentration?

d) [3 points] You now sequence the pool of barcodes obtained from the cell population before and after exposure to high temperatures. How can you tell which genes are important for cell survival under these conditions? Are you able to say something about the relative importance of these genes?

Solutions:

a) step 1: guess a number N larger than 5000, and pick N numbers (pseudo-)randomly from 1 to 5000, where N is the number of transformants (numbers) that are obtained and 5000 is the library size here.
step 2: count how many times each unique number (1 thru 5000) appears in that list of N randomly-chosen numbers.
step 3: find how many unique numbers (1 thru 5000) appear at least once.
step 4: take this as a ratio to all possible numbers (5000) to calculate the representation of each library member at least once.
step 5: adjust your guess of N to approach 99%

In MATLAB:

```
r = randi([1,5000],2.35e4,1);
for i = 1:1:5000
        n(i) = nnz(r == i);
end
nnz(n>=1)/5000
```

Where 2.35e4 = N = number of transformants needed to ensure 99% of the library members are represented at least once in the pool of transformants. This value gives 99.00% representation.

[Multiple approaches accepted including non-computational methods. For computational methods, either counting represented strains up to a certain N cells or the algorithm above with guess and check were accepted. -1/4 points for 4999 or 5001 strains due to coding errors, e.g. specifying strain identifier vector as 1 through 5001 when random integer generator is a,b inclusive, etc
-2 points for failing to show how the number was arrived at; -1 point for describing approach but not showing the code]

b) 5000 library members * 1000 copies = 5e6 DNA pieces
5e6 DNA molecules/6.022e23 molecules per mole = 8.3e-18 mol DNA
8.3e18 mol/50e-6 = 1.66e-13 M DNA = 0.166 pM DNA

[-1 point if strains used instead of cells
-1/2 point if units not in molarity]

c) answer from part b * 2^k = final DNA concentration < 1 nM, solve for k:
k = $\log_2$ (1 nM / 0.166 pM) = $\log_2$ (1000 pM / 0.166 pM) = 12.6 cycles ~ 12 cycles of PCR.

[-1 point for incorrect approach, e.g. wrong PCR formula, not accounting for volume or mismatched units; -1/2 point if decided to exclude template DNA from PCR formula; -1/2 point for rounding up]

d) [1 point] By sequencing and comparing the pool of barcodes that you end up with compared to the pool of barcodes that you began with, you are able to identify which barcodes were lost [1 point] and therefore which genes are essential for survival under these conditions. [1 point] The relative abundance before and after the experiment indicates the importance of each of those genes, with very important genes being completely absent in the final pool (ratio of 0) and unimportant genes being equally abundant in both pools (ratio of 1).

[-1/2 point for unspecific answers, e.g. not being clear how importance of a gene correlates to barcode prevalence]
[-1 point for not mentioning/discussing relative abundance with regards to quantifying barcodes]
[-2 points if quantification not discussed, e.g. “gene is important if it can help yeast survive”]
[no points deducted if barcode count per strain was discussed without normalizing to total barcodes, e.g. barcodes will stay the same for heat-sensitive strains while heat-insensitive strains will grow and have more abundant barcodes at the end]