Doug Chang

1) Starting with 10m copies of double stranded DNA, amplify a 1000 gene segment in each dna strand.

a) Assume the DNA polymerase makes no mistakes. Derive a formula for the total number of PCR amplicons that are produced from n copies of template DNA after k PCR cycles. Eacy cycle doubles the number of DNA strands. 2^k. $N2^k$

b) For a single cycle what is the probability a completly correct copy of the gene is generated? For a single cycle Assuming the PCR error rate P(e)=μ; From the law of total probability, $P(error) + P(no\ error) = 1$. $P(correct) = (1-\mu)$;

c) Assuming $\mu = 10^{-4}$ in the presence of Mn^{2+}. Derive a formula for the total number of correct PCR amplicons that are produced from N copes of temlpate dsDNA after k PCR cycles.

k = 0; N amplicons where n increments up to k.
k=1; $N + Np$
k=2; $N + Np + Np + Np^2 = N + 2Np + Np^2$
k=3; $N + Np + Np + Np^2 + Np + Np^2 + Np^2 + Np^3 = N + 3Np + 3Np^2 + Np^3$

The coefficients for p follows the binomial coefficients

$N * \binom{n}{k}$ for the coefficients where N is the initial concentration and p is 1 for no error rate.

The above progression becomes: $N(1+p)^k$

Represent the fraction of mutated and nonmutated by p where p is $\mu_{mut} + \mu_{correct}$ or $\mu_{correct} = 1 - \mu_{mut}$

$$N_{total} = N_{correct} + N_{mutated} = N * \binom{n}{k} p = N * (1+p)^k$$

$$N_{total} = N\binom{n}{k}(\mu_{correct} + \mu_{mut})^k$$

If we expand out p with both $\mu_{correct}\ and\ \mu_{mut}$ we only want the good terms and the mu mutated terms cross multiply. $p = \mu_{correct}$
and $\mu_{correct} = 1 - 10^{-4}$

$$N_{good} = N(1 + \mu_{correct})^k$$

d) Formula for mutated PCR amplicons.

$$N_{total} = N_{correct} + N_{mutation}$$

$$N_{total} = N2^k$$

$$N_{correct} = (1 + \mu_{correct})^k$$

$N_{mut} = N2^k - N(1 + \mu_{correct})^k$ where $\mu_{correct} = 1 - \mu_{mut} = 1 - 10^{-4}$

e) percentage of mutated strands for k=1,5,35. Assume this is mutated/good or mutated/total?,
Doesn't really matter either way since 2 is almost same as 1.9999

$$\frac{N_{mut}}{N_{good}} = \frac{N2^k - N(1 + \mu correct)^k}{N(1 + \mu_{correct})^k}$$

$$\frac{2^k - (1 + \mu_{correct})^k}{(1 + \mu_{correct})^k}$$

k=1 $\mu_{correct} = .9999\ 1 + \mu_{correct} = 1.9999$

$\frac{N_{mut}}{N_{correct}} = \frac{2 - 1.9999}{1.9999}$ **=** $5 * 10^{-5}$**; percentags is 5*10*-7**

k=5

$\frac{N_{mut}}{N_{correct}} = \frac{2^5 - (1.9999)^5}{1.9999^5}$ **= .0002; percentage is 2*10*-6**

k=35

$\frac{N_{mut}}{N_{correct}} = \frac{2^{35} - (1.9999)^{35}}{1.9999^{35}}$ **= .002; percentage is 2*10^-5**

f) compare above for 10^-4 and 4.4*10^-7

k=1 $\mu_{correct} = .9999999 \; 1 + \mu_{correct} = 1.9999999$

$\frac{N_{mut}}{N_{correct}} = \frac{2 - 1.9999}{1.9999}$ **= 5*10-8; percentage is 5^10-10**

k=5

$\frac{N_{mut}}{N_{correct}} = \frac{2^5 - (1.9999)^5}{1.9999^5}$ **= 2.5*10^-7, percentage 2.5*10^-9**

k=35

$\frac{N_{mut}}{N_{correct}} = \frac{2^{35} - (1.9999)^{35}}{1.9999^{35}}$ **= 1.7*10-6; percentage is 1.7*10-8**

g) mutation rate is μ_{mut}^3 for 3 errors.

$\mu_{good} = 1 - (10^{-4})$

$$\frac{N_{mut}}{N_{total}} = 10 * 10^6 * \frac{2^{20} - (1 + \mu_{good})^{20}}{2^{20}}$$ = 1

2a)
The size of a tartigrade genome is 100Mbp. Can make 10k clones. How big should DNA fragments be for 99% probability that protein coding gene will be contained in the library. Fragment size>> gene so just want 99% prob you are reading the 100Mbp accurately.

$N = \frac{ln(1-P)}{ln(1-f)}$ **;** P = .99 N = 10k, find F where f = F/L and L = 100Mbp.

$$10^4 = \frac{ln(1-.00)}{ln(1-\frac{F}{10^8})}$$

$$ln(1-\frac{F}{10^8}) = ln(1-.00) * 10^{-4}$$

$$e^{ln(1-.00)*10^{-4}} = 1 - \frac{F}{10^8}$$

$$1 - e^{ln(1-.00)*10^{-4}} = \frac{F}{10^8}$$

$$10^8(1 - e^{ln(1-.00)*10^{-4}}) = F$$

F=100.5 or 101bp

b) If the gene length is ~ fragment/read length how does the above equation change.

Expected Reads within G , region of Gene length, is $\frac{N}{L} * G = Ng$

From the poisson distribution, P(no reads in Gene) = e^{-Ng}

P(no reads in Gene Interval for all N strands) = $(e^{Ng})^N$

P(at least one N clones contains GOI) + P(none of N clones/strands) = 1
1-P(at least 1 N clones contains GOI) = $e^{N^2 g}$

$$ln(1-P) = N^2 g ln(e)$$

$$N = \sqrt{(\frac{1-P}{g})}$$

c) Coverage is defined as $C = \frac{N * F}{L}$ where N is the number of clones, F is fragment size, L is genome size. For 10x coverage, where C = 10; what is the P(Nucleotide) is in the fragment pool.

$P(1readinF) = \frac{C^y e^{-C}}{y}$ where C is the coverage defined above.
$P(readinF) = \frac{10^1 e^{-}10}{1} = .45\%$

d) 2 ways to construct replacement for functional assays for the mystery protein.

1) RnaSEQ the dormant tartigrade and compare the cDNA vs. the gDNA. Assumes the rna is expressing the mystery protein when Mr. Tartigrade is dormant. This allows you to check the proteins created from the mRNA. Can also compare a dormant vs. NonDormant tartigrade, euthanize both of them and extract RNA from tissue and compare the different mRNA sequences to see if there is a difference indicating the dormant sample is expressing the protein as it hibernates. This is a comparison of 2 RNASeq samples.
2) If the mystery protein is not a funciton of environmental factors and is on both the dormant and nondormant samples, you can figure out the genes in the DNA by add point mutations and seeing which point mutations cause a sample to not go dormant as compared to a control which has no modifications. This is a DNA modification.

3) LIbrary prep.

3a)
#given a set of 5000 cells, if you randomly pick one, how many random picks do you need to get 99% of all 5000? **23849**

cut and paste from jupyter notebook

#given a set of 5000 cells, if you randomly pick one, how many random picks do you need to get 99% of all 5000?

```
import random
#4950
dict={}
N=0
random.seed(42)

while len(dict.keys()) < 4950:
    pick = random.randint(1,5001)
    if(N%1000==0):
        print(N)
    if pick not in dict:
        dict[pick]=1
    else:
        dict[pick] +=1
    N += 1
print (len(dict.keys()))
print(N)
```

Output:

```
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
17000
18000
19000
20000
21000
22000
23000
```

```
4950
23849
```

How to compute this into drop size? or amount of liquid? There must be some convention where the yeast is in some growth media liquid. Then you take a x microliters of this. convert use

b) 1000 copies of each barcode N_b. Run the above program with a modification to register 1k copies of each barcode. We need 23849 cells/dna strands to get 99% of 5k. Multiply 23859 by 1000 to get number of cells/strands of DNA.

$$\frac{1\,mole}{6*10^{23}} = \frac{x}{(23849)1*10^3}; \text{x=}4*10^{-17}\,moles$$

I have no idea how 50 microliters is used. Need more chemE experience! Assume some molar/liter units but the announcement on the hw said to answer in “moles”

c) We start with 23849 DNA molecules. We do PCR cycles to get to 1nM. 1nM = $(6*10^{23})(1*10^{-}9) = 6*10^{14}$.

$$23849*(2^x) = 6*10^{14}$$

$$2^x = \frac{6*10^{14}}{23849}$$

$$xln(2) = ln(\frac{6*10^{14}}{23849})$$

x=35

d) Sequence the barcode genes before and after high temperature exposure. If there is a control group where the population statistics can be compared to the group which is exposed to high temperature there may be a difference in gene expression. This also makes it difficult to guarantee reproducbility if there are other experimental factors influencing population under heat.