Name: KEY
Recitation Section:

7.03 Problem Set 5

Due Monday, April 27, 2015 by 3 PM

Turn into the box outside the BioEd office in building 68

Remember to show your work for all questions.

Corresponding p-values for χ^2:

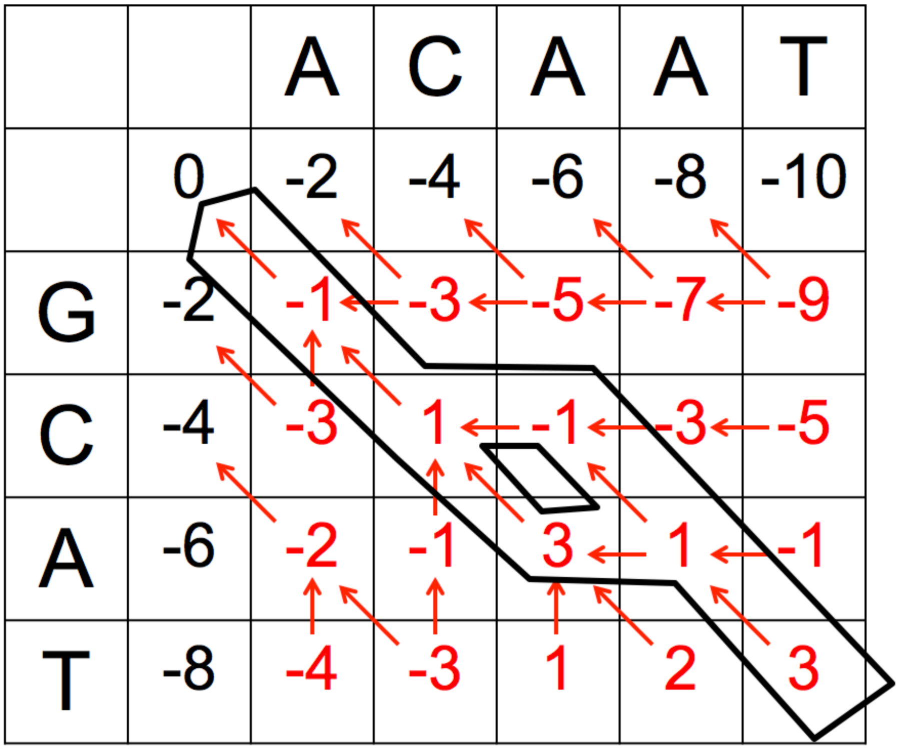
p-value	**.995**	**.975**	**0.9**	**0.5**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**
df = 1	.000	.000	.016	.46	2.7	3.8	5.0	6.6	7.9
df = 2	.01	.05	.21	1.4	4.6	6.0	7.4	9.2	10.6
df = 3	.07	.22	.58	2.4	6.3	7.8	9.3	11.3	12.8

Question 1:[10 points]

You have recently sequenced and assembled the genome of a recently discovered species of fruit fly, *Drosophila fictionalis*. You decide that, in order to learn more about its genome, you will compare it to two other commonly studied species, *D. melanogaster* and *D. simulans*.

You first choose a small region of the *D. fictionalis* genome that is predicted to code for protein because it contains an ORF. You decide to start by doing a **pairwise alignment** to a similar region of the *D. melanogaster* genome.

a) In the matrix below, fill the matrix and indicate the traceback path for a **global alignment** using the Needleman-Wunsch algorithm. Use a match score of **m = 2**, a mismatch penalty of **s = -1**, and a gap penalty of **d = -2**. The marginal scores have already been filled in for you.



		A	C	A	A	T
	0	-2	-4	-6	-8	-10
G	-2	-1	-3	-5	-7	-9
C	-4	-3	1	-1	-3	-5
A	-6	-2	-1	3	1	-1
T	-8	-4	-3	1	2	3

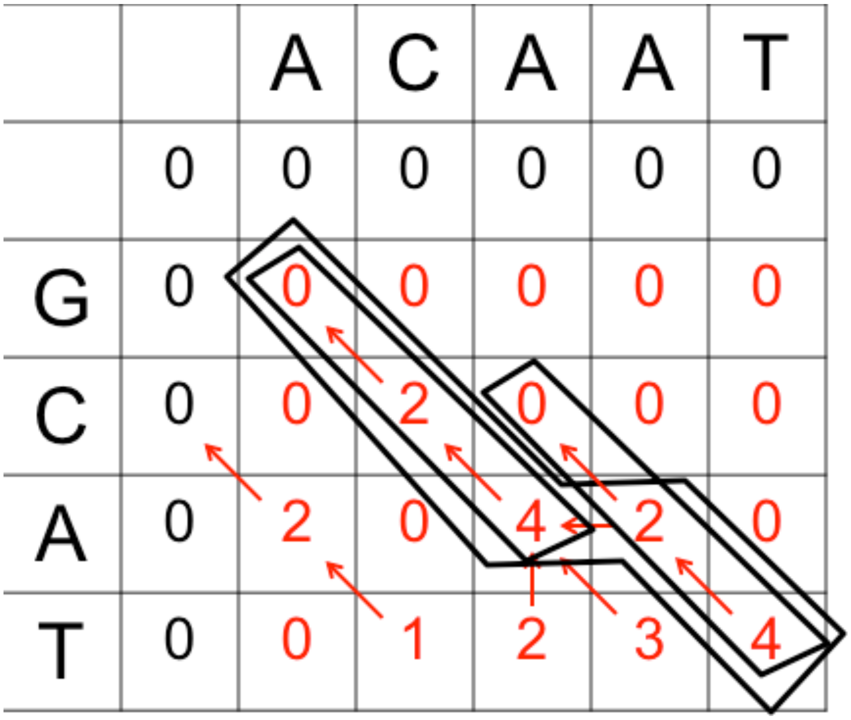
[1 point]

b) Write the optimal global alignment found using this algorithm. (If there is more than one optimal alignment, write them all.)

ACAAT or ACAAT
GCA-T GC-AT

[1 point, 0.5 for each alignment]

c) In the matrix below, fill the matrix and indicate the traceback path for a **local alignment** using the Smith-Waterman algorithm. Use a match score of **m = 2**, a mismatch penalty of **s = -1**, and a gap penalty of **d = -2**. The marginal scores have already been filled in for you.



		A	C	A	A	T
	0	0	0	0	0	0
G	0	0	0	0	0	0
C	0	0	2	0	0	0
A	0	2	0	4	2	0
T	0	0	1	2	3	4

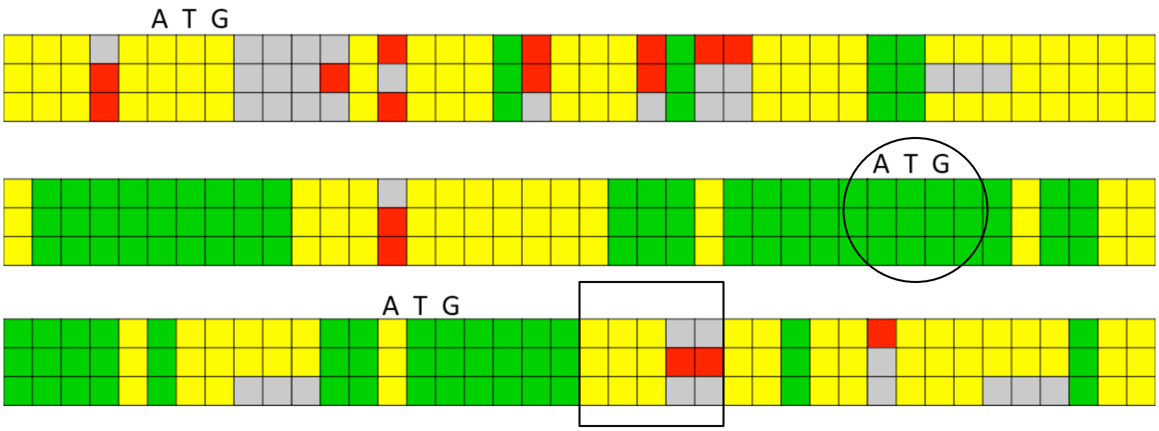
[1 point]

d) Write the top-scoring local alignment found using this algorithm. (If there is more than one top-scoring alignment, write them all.)

```
CA      and     CAAT   and     AT
CA              CA-T           AT
```

[1 point; 0.5 for each alignment]

You decide that this pairwise alignment isn't giving you enough information, so you do a **multiple alignment** of the region between *D. fictionalis*, *D. melanogaster*, and *D. simulans*. The result of that alignment is shown below:



e) There are three start codons within this region (shown by the letters "ATG" above three consecutive bases in the diagram). On the diagram, circle the one that you think is the actual translation start site for a protein, and explain your reason for choosing that one below.

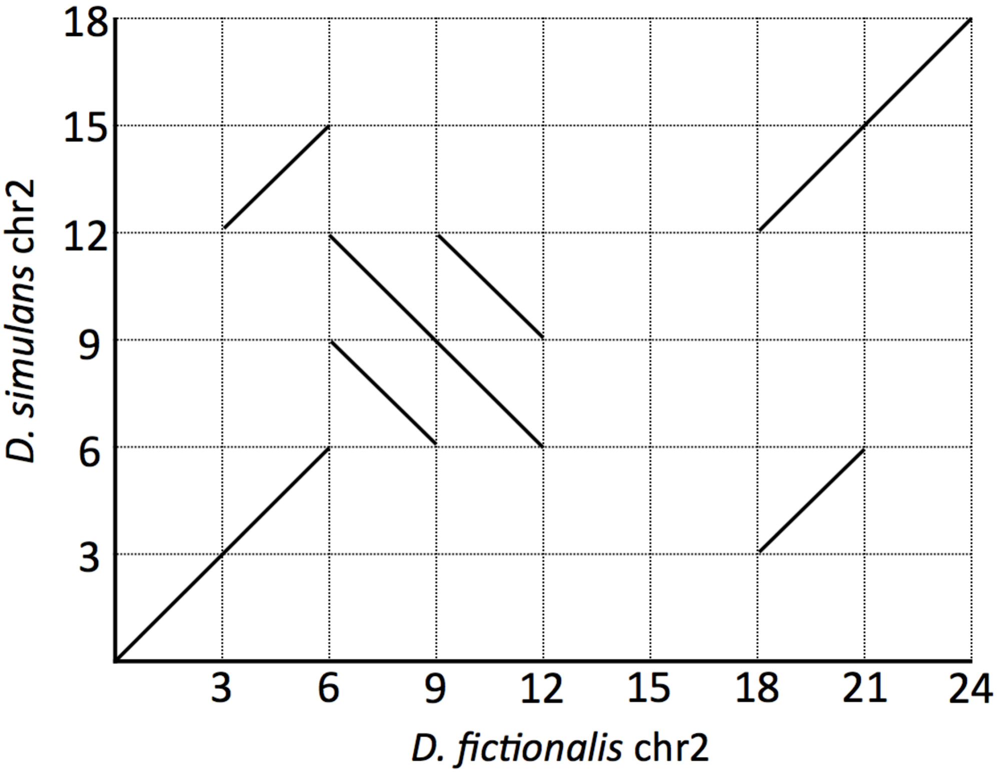
That is the only ATG that is conserved and that has a stretch of high homology without any frameshifts following it.

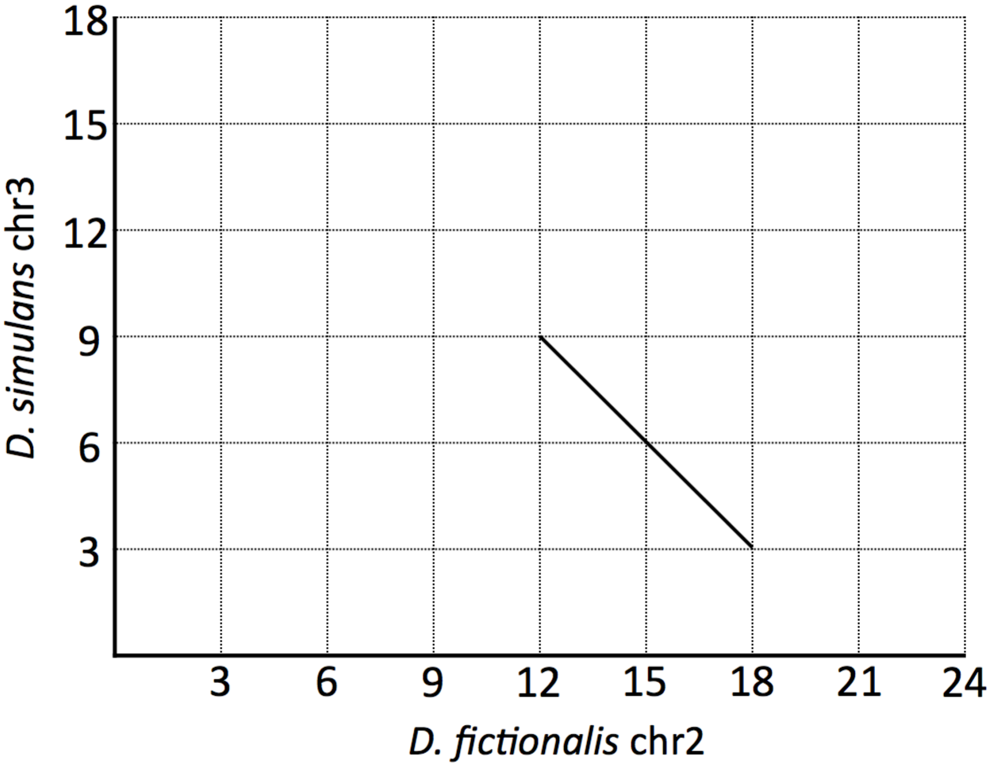
[1 point; 0.5 for correct location, 0.5 for explanation]

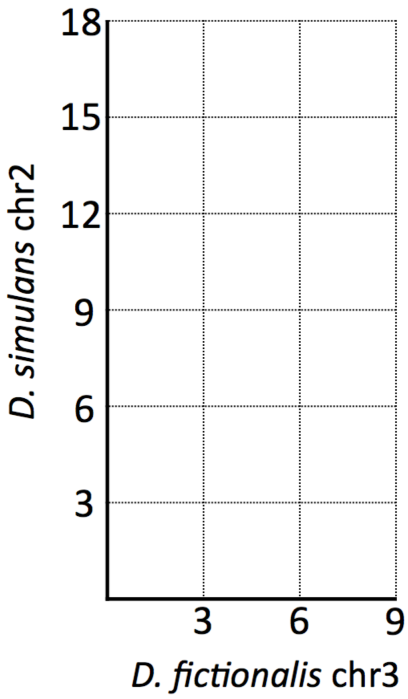
f) After the translation start site, there seems to be a point where the high level of homology stops, and the sequence no longer appears to code for protein. However, there are no stop codons before then. Put a square around the (approximate) point where the protein coding region ends, and explain below how this could happen without a stop codon.

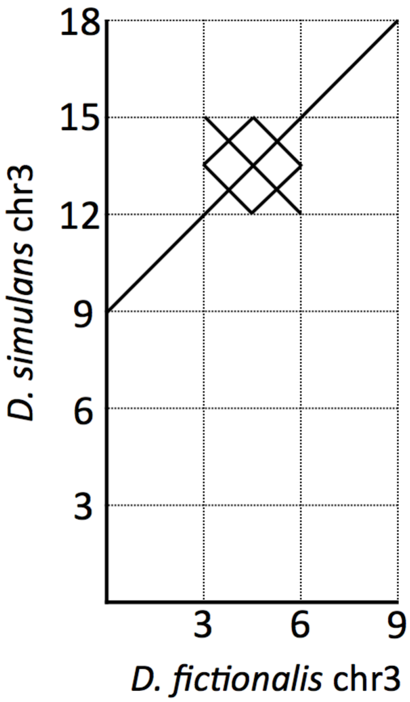
Since these sequences are in flies, there must be a splice site near this point.
[1 point; 0.5 for correct location, 0.5 for explanation]

After comparing individual genes, you decide that you want to compare these species on a genome-wide scale. You compare chromosomes 2 and 3 of *D. fictionalis* with chromosomes 2 and 3 of *D. simulans* by making dot plots of them. The results are shown below (axis labels are in megabases):

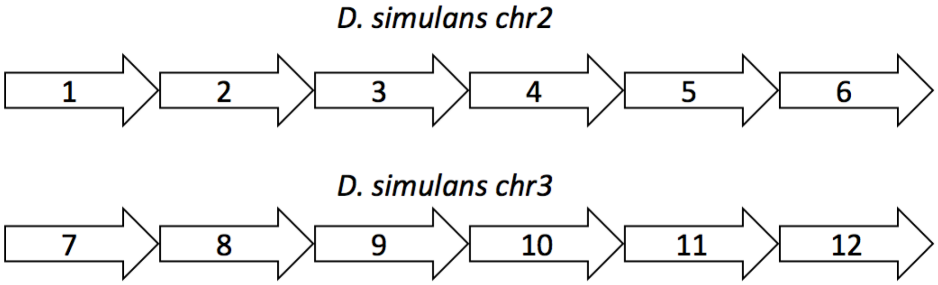

D. simulans chr2
18
15
12
9
6
3
3
6
9
12
15
18
21
24
D. fictionalis chr2


D. simulans chr3
18
15
12
9
6
3
3
6
9
12
15
18
21
24
D. fictionalis chr2

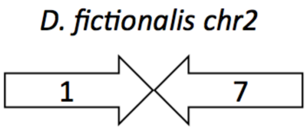




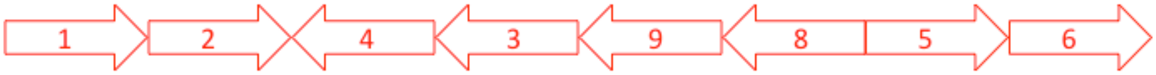
g) Chromosomes 2 and 3 of *D. simulans* can be drawn as **3 Mb-long blocks**, numbered from 1-12, as shown below:



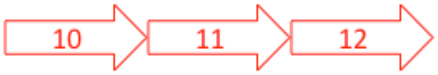
Draw chromosomes 2 and 3 of *D. fictionalis* using the blocks of the *D. simulans* chromosomes shown above, including the correct orientation. **Do not draw any structural elements that occur *within* single 3 Mb-long regions.** For example, if the first 3 Mb of chr2 of *D. fictionalis* is homologous to the first 3 Mb of *D. simulans* chr2, and the next 3 Mb of chr2 of *D. fictionalis* is homologous to the first 3 Mb of *D. simulans* **chr3** but in the opposite orientation, you would draw:



D. fictionalis chr2:



D. fictionalis chr3:



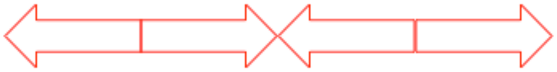
[2 points; 1 point for each chromosome]

h) Are there any 3 Mb-long regions **within** the *D. fictionalis* genome that are identical to each other? List their coordinates **in the *D. fictionalis* genome**. What is their orientation—direct or inverted?

Mb 3-6 and 18-21 of chr2 are direct repeats. Mb 6-9 and 9-12 of chr2 are also direct repeats.

[1 point; 0.5 points for each region of homology]

i) From the dot plots, it is clear that there is some kind of complex internal structure from Mb 3-6 on chromosome 3 of *D. fictionalis* (and on its homologous region in *D simulans*.) Draw that structure below, using arrows to represent identical regions and their orientation.



[1 point]

Question 2: [7 points]

Immunoglobulin E (IgE) is a type of antibody that functions to fight parasitic helminth infections. One of the domains that makes up and IgE antibody, Cε3, is found in two co-dominant alleles in humans, which we will designate A and B. The B allele is known to be much less effective at fighting off parasites. You decide to examine these variants in a human population genetics study.

You first wish to determine whether the two alleles exist in Hardy Weinberg Equilibrium.

a) State 5 of the assumptions that the Hardy Weinberg model makes. (1 point)

1. Mating is random
2. No new mutations arise
3. No selection is present
4. No genetic drift/founder effect
5. No migration

0.2 each for any valid assumption

b) You start your study in America, where helminth parasites are not much of a public health concern due to eradication of insect vectors, safe disposal practices of waste and the availability of drugs. After genotyping 1000 American individuals for the A and B alleles, you obtain the following results.

A/A	A/B	B/B	p	q
679	287	34	0.8225	0.1775

(i) What are the allele frequencies, *p* (frequency of A) and *q* (frequency of B)? Add them to the table above and please show your work. (1 point)

p= [679+ (287/2)]/1000 = 0.8225
q= [34+ (287/2)]/1000 = 0.1775

no credit if used p^2= [A/A]/Total, we do not know if we are in HWE

(ii) Is this population in Hardy Weinberg Equilibrium? Please show your work. (2 points)

Need to perform a chi-squared test.

H_0= Population in HWE
H_A= Population not in HWE

Expected frequencies:

A/A	A/B	B/B
$0.8225^2*1000=$ 676.5	$2*0.8225*0.1775*1000=$ 291.9	$0.1775^2*1000=$ 31.5

Genotype	**Observed**	**Expected**	**$(O-E)^2$**	**$(O-E)^2/E$**
A/A	679	676.5	6.25	0.009
A/B	287	291.9	24.01	0.08
B/B	34	31.5	6.25	0.2

$\Sigma(O-E)^2/E = 0.289$

Deg. of freedom = #classes - # estimated parameters = 3-1-1 = 1

Use the χ^2 table to find p-value:

p-value	**.995**	**.975**	**0.9**	**0.5**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**
df = 1	.000	.000	.016	.46	2.7	3.8	5.0	6.6	7.9
df = 2	.01	.05	.21	1.4	4.6	6.0	7.4	9.2	10.6
df = 3	.07	.22	.58	2.4	6.3	7.8	9.3	11.3	12.8

$p > 0.5$

We accept the null hypothesis; this population is in Hardy Weinberg Equilibrium.

½ point for correct test

½ point for right conclusion

1 point for the right math (within rounding error)

c) You repeat your experiment and genotype 1000 individuals for the A and B alleles in a country where parasitic helminthes are still a major public health problem. Based on this information, state which of the assumptions described above does not apply to this population? (0.5 points)

Selection is acting against the B allele

d) Suppose the selective disadvantage of the B allele = 0.6. Given the following experimental results and assuming the population is at steady state, what is the mutation rate, μ, of the B allele? Does this result indicate the population is at steady state for the A and B alleles? (1 point)

A/A	A/B	B/B	p	q
810	180	10	0.9	0.1

In steady state, $\Delta q_{sel} + \Delta q_{mut} = 0$, so $-Sq^2+\mu=0$, so $\mu=Sq^2$

$\mu= 0.6*0.1^2 = 0.006$

No, this is a very high mutation rate, so we expect that the B allele frequency is still in decline.

½ point for correct mutation rate (also gave points for $p\mu=Sq^2$ so $\mu= 0.0067$
½ point for conclusion

e) Assuming this population is NOT in steady state (allele frequencies are still changing), estimate the genotype frequency of B/B and the new allele frequencies, p' and q', in the next generation. (1 point)

After selection: $f(B/B) = q^2(1-S) = 0.1^2(0.4) = 0.004$
This reflects the B/B individuals that will reproduce.

$q' = (0.004/0.994) + ½ *(0.18/0.994) = 0.094$
$p' = 1-q' = 0.906$

new f(B/B) in gen 2 = $q'^2= 0.0088$

Alternatively, can calculate $\Delta q_{sel}= -Sq^2 = -(0.6)*(0.1^2) = -0.006$
$q'= q- \Delta q_{sel} = 0.1-0.006 = 0.094$

¼ point using relevant equation
¼ point for finding f(B/B) after selection or in next generation (unclear working)
¼ ea for p' and q'

f) You decide to study yet another population which also suffers greatly from infectious helminth diseases (same selective disadvantage cost as the previous population), but who live on an island where a certain species of flower causes severe allergic reactions in A/A individuals. It has become socially undesirable to have these allergies, and so A/A individuals are at a fitness disadvantage compared to A/B individuals. Without doing any calculations, would you expect the B allele frequencies to display similar changes to the population considered in parts c-e? Justify your answer **briefly.** (0.5 points)

No. Here there is a heterozygote advantage acting to balance out selection against the B allele, so it will remain present at a higher level than in the previous population even after reaching steady state.

Question 3: [3 points]

You are interested studying Greig cephalopolydyndactyly syndrome (GCPS), a disorder that affects development of the limbs, head, and face. The gene responsible for GCPS in humans is GLI3 that has a homolog in mice, Gli3. For convenience, you decide to study the gene in field mice. Mutations in this gene are in an autosomal recessive manner, which we can designate as *gli3-*. Many functional alleles of Gli3 exist in wild mice populations ($Gli3^a$, $Gli3^b$, $Gli3^c$, etc) and the frequency of the only disease causing allele (*gli3-*) is very rare, $2*10^{-7}$.

a) What is the probability that two unrelated wild mice will have a pup born with GCPS? (0.5 points)

$p(dd) = (2*10^{-7})^2 = 4*10^{-14}$

no credit given for $(1/4)q^2$

b) You collect a couple of mice and start to inbreed them. What is the probability that mating between the equivalent of two 1st cousins will result in a pup born with GPCS? Between a brother-sister mating? Assume the original mice carry all different alleles. (1 point)

p(dd) = p(homozygous by descent) x p(allele is d)

For the cousins:
p(homozygous by descent) = $F = [(1/2)^2 (1/2)^2*(1/4)]*4 = 1/16$

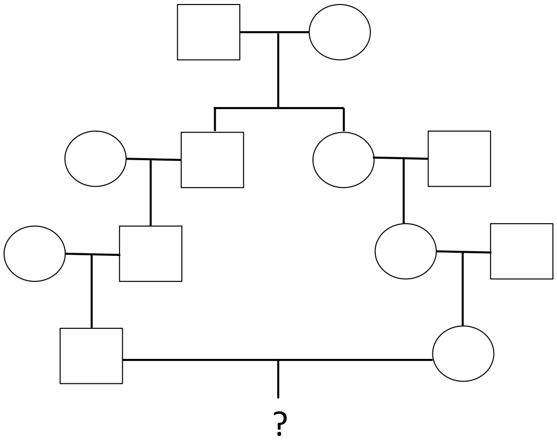
$p(d) = (1/16)*(2*10^{-7}) = \mathbf{1.25*10^{-8}}$

For brother/sister:
p(homozygous by descent) = $F = [(1/2)*(1/2)*(1/4)]*4$ = ¼

$p(d) = (1/4)*(2*10^{-7}) = \mathbf{5*10^{-8}}$

½ point each correct answer, ½ credit given if did not multiply by 4 since we could have this happen at any of the alleles.
-¼ for minor math error

c) What is the probability that mating between the two related mice indicated in the pedigree below will result in a pup born with GCPS? Assume the first generation carry all different alleles. (1 point)



p(dd) = p(homozygous by descent) x p(allele is d)

p(homozygous by descent) = F = $[(1/2)^3 (1/2)^3*(1/4)]*4 = 1/64$

p(d) = $(1/64)*(2*10^{-7})$ = **$3.125*10^{-9}$**

½ point for F, ½ point for correct answer
½ credit given if did not multiply by 4 since we could have this happen at any of the alleles.
-¼ for minor math error

d) Considering that there are approximately 20,000 genes in the mouse genome, in how many genes would the child indicated by the “?” be expected to be homozygous?
(0.5 points)

Number of homozygous genes = F * number of total genes

= (1/64) * 20,000 = **312.5 homozygous genes**

No penalty if carried over incorrect F