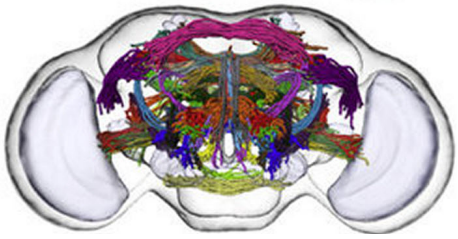
Bio393: Genetic Analysis **Name:**_____
Problem Set #4
Due on Friday, June 2, 3 PM

Question 1:
Drosophila extend their proboscis to eat or drink. This behavior is controlled by a complex system of connected neurons and muscles. The four images below show the process of proboscis extension (temporally from left to right).



(a) You identified a mutant with a recessive phenotype of much reduced proboscis extension that you named *hungry*, and you are interested to figure out in which set of neurons shown in different colors (left) the gene regulates proboscis extension. Using a series of promoters that express gene products in these regions, please describe how you would determine in which set of neurons the *hungry* gene acts. Also, describe any caveats.

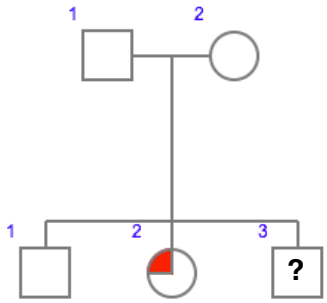
You can rescue the mutant hungry phenotype using the neuronal-specific promoters driving expression of the wild-type hungry gene. If hungry acts in that neuron group, then the mutant reduced proboscis extension phenotype will be rescued and flies will extend their proboscises. The major caveat is over expression or misexpression could rescue the phenotype by bypassing the requirement of hungry in another neuron type.

(b) You determine that *hungry* acts in the purple set of neurons to regulate proboscis extension. Using the same promoter and optogenetic tools, describe how you would use specific tools to determine if those neurons inhibit or promote proboscis extension.

Using the same neuronal promoter that rescues the hungry mutant phenotype, you can express channelrhodopsin (activating) or halorhodopsin (inhibiting) in wild-type flies in those neurons. Upon stimulation with the appropriate wavelength of light, the neuron will be activated or inhibited. You can measure the proboscis extension. If the neuron is activating, then channelrhodopsin will stimulate extension and halorhodopsin will inhibit extension. If the neuron is inhibitory, then channelrhodopsin will inhibit extension and halorhodopsin will stimulate extension.

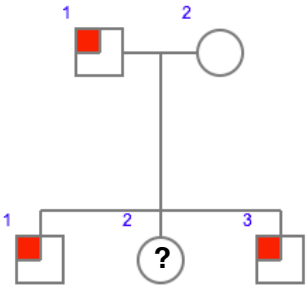
Question 2:
Each of the families below exhibits a different, extremely rare genetic disorder. Individuals expressing the trait (the disorder) are indicated by symbols with red sections. Assume that no new mutations have arisen in any of the individuals shown. Consider the following possible modes of inheritance: (i) X-linked recessive with complete penetrance, (ii) autosomal recessive with complete penetrance, (iii) autosomal recessive with 70% penetrance, (iv) autosomal dominant with complete penetrance, (v) autosomal dominant with 70% penetrance. For each pedigree state which, if any, of these five modes of inheritance are not possible. For the modes of inheritance that are possible, calculate the probability that the individual indicated by a "?" is affected.

(a)



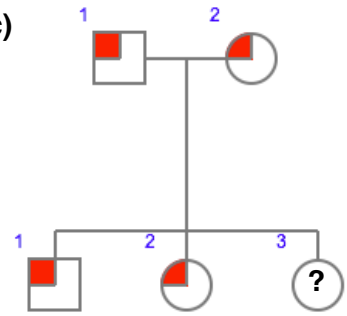
i. Not possible because individual I-1 is not affected.
ii. Possible, individual II-3 has a 1/4 chance of being affected or 1/2 of being a carrier
*iii. Possible, individual II-3 has a 1/4 * 7/10 chance of being affected or 7/40.*
iv. Not possible, neither parent is affected
*v. Possible, one parent must be a carrier but not fully penetrant. The probability that individual II-3 will be affected is 1/2 * 7/10 = 7/20*

(b)



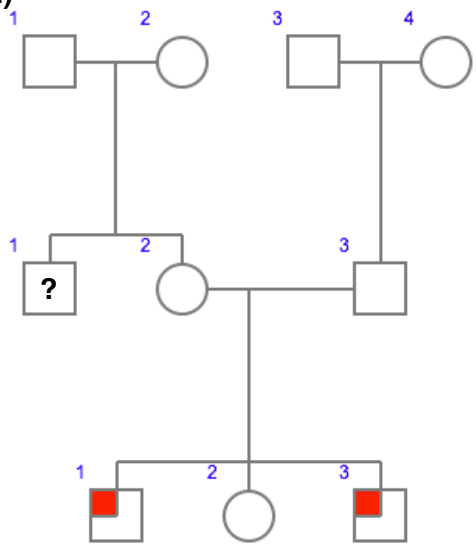
i. Possible if individual I-2 is a carrier. II-2 has a 1/2 chance of being affected
ii. Possible, same as i, 1/2 chance of being affected
*iii. Possible, because of penetrance the chance is 1/2 * 7/10 = 7/20*
iv. Possible, 1/2 chance of individual II-2 being affected
*v. Possible, same logic as iv except penetrance changes the chance to 1/2 * 7/10 = 7/20*

(c)



i. Possible, individual II-3 has 100% chance of being affected
ii. Possible, individual II-3 has 100% chance of being affected
iii. Possible, individual II-3 has 100% chance of inheriting the affected allele but a 7/10 chance of being fully penetrant so 7/10 chance overall
iv. Possible, because the disease is rare both parents are likely heterozygotes, individual II-3 has a 3/4 chance of being affected
*v. Possible, same logic as iv, because the disease is rare both parents are likely heterozygotes, individual II-3 has a 3/4 chance of being affected, but penetrance makes the chance 3/4 * 7/10 = 21/40*

(d)



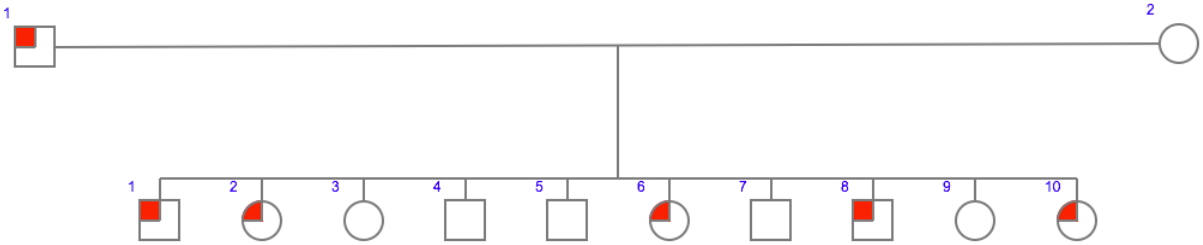
i. Possible, individual II-1 has 1/2 chance of being affected
ii. Possible, but pretty unlikely both parents would have to be carriers. For a rare disease, it is highly unlikely.
iii. Possible, but pretty unlikely both parents would have to be carriers. For a rare disease, it is highly unlikely.
iv. Not possible, no affected individuals in generations I or II
*v. Possible, individuals in generations I or II would have to be not fully penetrant. Individual I-1 or I-2 could be a carrier and not penetrant for the disease. Because it is a rare disease, we expect that both would not be carriers. If one is a carrier, then II-1 has a 1/2 * 7/10 or 7/20 chance of expressing the disease.*

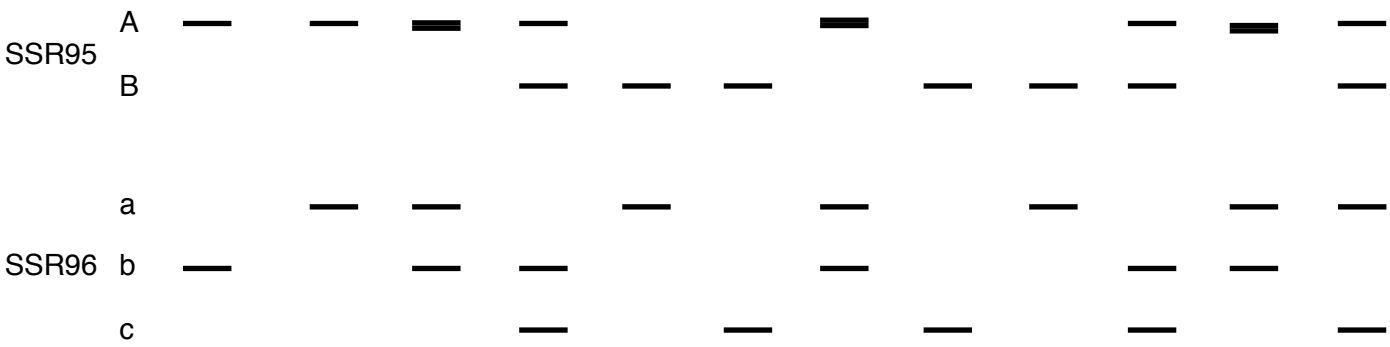
*However, we also do not know what side of the pedigree the allele conferring the dominant phenotype would come from, so we need to take into account that probability (1/2). 1/2 * 1/2 * 7/10 = 7/40*

Question 3:
You set out to genetically map color blindness with respect to short-sequence repeat (SSR) markers. Color blindness shows X-linked recessive inheritance and therefore is usually found in males. However, the mutant allele frequency is sufficiently high that colorblind females do occur.

Alleles: + (normal) cb (associated with color blindness)

Here is a family in which some individuals are affected:





(a) Diagram the two possible phase relationships between the SSR95 and SSR96 alleles in the mother.

$$\frac{A\ a}{B\ c} \qquad \frac{A\ c}{B\ a}$$

(b) Calculate the LOD score for linkage at θ = 0.1 between SSR95 and SSR96 in this family.

$$LOD_{0.1} = \log_{10} \frac{1/2 * (0.9)^8 * (0.1)^2 + 1/2 * (0.9)^2 * (0.1)^8}{(1/2)^{10}} = 0.343$$

(c) Identify a value of θ at which this family will yield a higher LOD score for linkage between SSR95 and SSR96. Calculate the LOD score for linkage between SSR95 and SSR96 at that new θ value.

Two out of 10 recombinant suggests that a theta of 0.2 would be better

$$LOD_{0.2} = \log_{10} \frac{1/2 * (0.8)^8 * (0.2)^2 + 1/2 * (0.8)^2 * (0.2)^8}{(1/2)^{10}} = 0.536$$

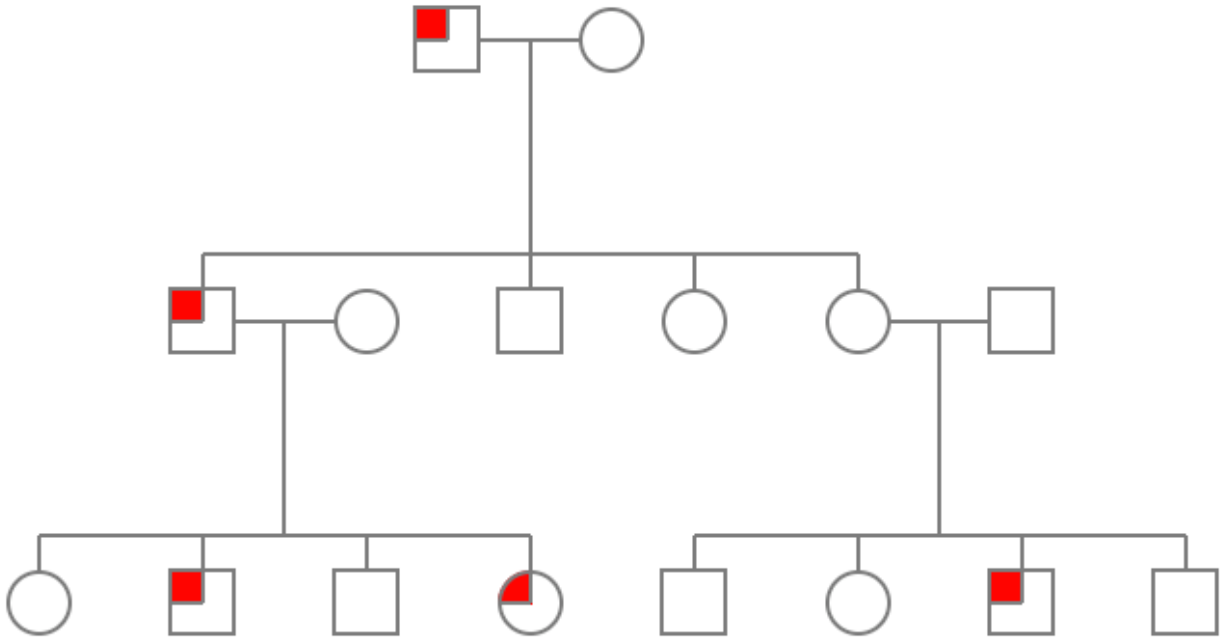
(d) Diagram the two possible phase relationships between the SSR95 and color blindness alleles in the mother.

$$\frac{A\ cb}{B\ +} \qquad \frac{A\ +}{B\ cb}$$

(e) Calculate a LOD score for linkage at θ = 0.1 between SSR95 and color blindness in this family.

$$LOD_{0.1} = \log_{10} \frac{1/2 * (0.9)^9 * (0.1)^1 + 1/2 * (0.9)^1 * (0.1)^9}{(1/2)^{10}} = 1.297$$

Question 4:
Syndactyly is a rare genetic condition inherited as an autosomal dominant trait. Unusually however, a person who has the defective allele responsible for syndactyly (N) does not always express the trait. The diagram below shows a pedigree of a family with syndactyly.



You find no history of syndactyly in the ancestors of individual I-2, II-2, and II-6. Assuming no new mutations exist, explain why or why not it is reasonable to conclude the following:

(a) Individual II-5 has the genotype nn at the syndactyly locus

It is not reasonable to conclude that individual II-5 is nn because she has offspring that have the genotype Nn, and (1) II-6 has no ancestors with syndactyly and (2) no new mutations are arising.

(b) Individuals II-4 and II-5 may have the same genotype at the syndactyly locus

It is reasonable that individual II-4 and II-5 both have the same genotype. Because II-4 has no offspring and both individuals do express the syndactyly trait, it is possible that they have different genotypes. However, we believe that II-5 has the genoypte Nn, and there is a 50% chance that II-4 has the same genotype.

(c) Individual II-2 has the genotype Nn at the syndactyly locus

It is not reasonable to conclude that II-2 is a carrier for syndactyly and not expressing because we see no family history of syndactyly for her and the affected offspring likely come from the father for this rare disorder.

(d) Individuals III-2 and III-7 have different genotypes at the syndactyly locus

Given that (1) syndactyly is a rare disorder, (2) II-2 and II-6 have no family history of the disorder, and (3) no new mutations are arising to cause syndactyly, it is likely that III-2 and III-7 both have the same genotype (Nn).

Question 5:
You are studying a dominant Mendelian disease via linkage analysis and are focusing on a single marker. Two large families have been genotyped at the same marker and scored for the disease.

In Family I, ten offspring are genotyped: eight children inherited a marker allele and a disease-causing allele without recombination; two children appear to be recombinants. You test many values of the recombination fraction (theta) and discover that theta = 0.2 gives the maximum odds ratio, which is 6.87 (LOD = 0.837).

In Family II, 20 offspring are genotyped: 17 children inherited a marker allele and a disease-causing allele without recombination; three children appear to be recombinants. You test many values of theta and discover that theta = 0.15 gives the maximum odds ratio, which is 223.4 (LOD = 2.34).

To combine data across Family I and Family II, you multiply odds ratios (add LOD scores). The final estimate of the odds of linkage relative to the null as 1534.8 (LOD = 3.18). Explain what is wrong with this calculation.

You can not combine LOD scores when they are calculated using different values of theta.

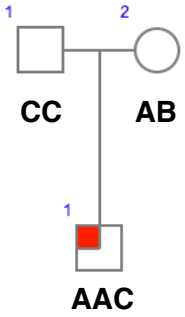
Question 6:
Imagine you are doing a genome-wide linkage study in Finnish families looking for the genetic determinants of blood pressure in humans. You have five multi-generational families; each individual is genotyped at 1000 markers and his/her blood pressure is measured. A recent, published study in Icelandic families identified a highly significant locus on chromosome 10 responsible for blood pressure variation. You look through your results and see no significant linkage between the genotype and the disease in your data. Your nearest marker to this locus is 30 cM away.

Give three reasons why you might have failed to find linkage to the chromosome 10 locus. Please explain each reason with no more than one to two sentences.

(1) Locus heterogeneity: the chromosome 10 locus is not polymorphic in the Finnish families and not responsible for disease variation.
(2) Complexity: the chromosome 10 locus is polymorphic in your family but there are other causative loci with greater effect.
(3) Marker spacing: your closest marker is too far away to give an odds of linkage that can be distinguished from the null hypothesis.
(4) Number of patients: your number of individuals in the Finnish families is too small to get significant linkage.
(5) Number of markers: because of multiple testing of 1000 markers, your significance threshold is too high to detect linkage
(6) Environment: the individuals in your families have very different lifestyle habits and this environmental variation trumps modest genetic effects.

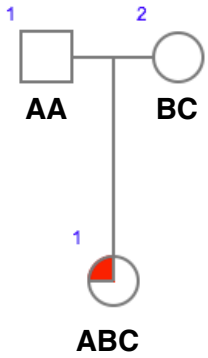
Question 7:
Chromosomal abnormalities cause a large fraction of aborted pregnancies and severe developmental disorders. For the following two parts, please describe what caused the inheritance of the extra chromosome in the affected child?

(a) Down syndrome is caused by inheritance of an extra copy of chromosome 21. A marker on chromosome 21 was genotyped in both parents and child shown below.



The extra chromosome was caused by non-disjunction during meiosis II. Sister chromatids harboring the A allele of the chromosome 21 marker failed to go to separate gametes.

(b) Patau syndrome is caused by inheritance of an extra copy of chromosome 13. A marker on chromosome 13 was genotyped in both parents and child shown below.



The extra chromosome was caused by non-disjunction during meiosis I. Homologous chromosomes failed to go to separate gametes, leading to both the A and B alleles of the chromosome 13 marker being inherited.

Question 8:
You are studying a rare recessive disease that you have mapped approximately by linkage to simple sequence repeat (SSR) markers. In an effort to localize the disease locus more precisely, you decide to look for linkage disequilibrium (LD) with respect to two dimorphic DNA-based markers (designated A and B) known to be in the vicinity of the disease gene. You first examine a relatively isolated Scandinavian population in which the frequencies of alleles A1 and A2 are 0.9 and 0.1 respectively, and the frequencies of B1 and B2 are both 0.5. By examining the DNA from individuals in the population who have the disease it is possible to determine the frequency of each haplotype, as shown in the table below.

Haplotype	Number of individuals with the disease
A1 B1	10
A1 B2	90
A2 B1	1
A2 B2	10

(a) (i) What can you say about possible linkage disequilibrium between each of the markers and the disease causing allele in this population? (ii) Assume the disease causing allele arose after both of the markers (A and B) were present in the population. Which of the two DNA-based markers is likely to be closer to the disease locus? (iii) Assuming that the disease allele arose only once in this population, what can you say about the haplotype context in which the original disease mutation arose?

i. We do not know the number of individuals with the different haplotypes that do NOT have the disease, so we can't calculate D for the disease and either marker. However, we can compare the allele frequencies in the general population to those in the affected population.

For marker A:
Ratio of A1:A2 in the general population is 9:1 and (10+90 : 1+10) or 9.1:1 in the affected population. These two values are close so marker is likely in linkage equilibrium.

For marker B:
Ratio of B1:B2 in the general population is 1:1 and (10+1 : 90+10) or 1:9.1 in the affected population. These two values are different so marker is likely in linkage disequilibrium.

ii. For the reasons in i, marker B is likely closer

iii. It is likely that the disease-causing allele arose in the B2 haplotype background because more affected individuals have the B2 allele. We can't tell which A allele was present at the time the disease-causing allele arose. There has been too much recombination and the A marker is in equilibrium with the disease-causing allele.

(b) Next you examine the genotypes of individuals with the same disease in a large African population. In this population the frequencies of alleles A1 and A2 are both 0.5, and the frequencies of B1 and B2 are also both 0.5. The frequencies of the each haplotype for individuals with the disease in the African population are shown in the table below.

Haplotype	Number of individuals with the disease
A1 B1	26
A1 B2	24
A2 B1	28
A2 B2	22

Give two different explanations for why the linkage disequilibrium results differ between the African and Scandinavian populations.

The two markers appear to be in linkage equilibrium with the disease-causing allele.

Some explanations are:

1. *Because the African population is older than the Scandinavian population, it is possible that more recombination occurred between the disease-causing allele and the B marker.*
2. *The disease-causing allele could have arisen independently multiple times in the B1 and B2 haplotypes.*
3. *The disease in the African population is caused by mutation in a different gene unlinked from the A and B markers.*
4. *One explanation might be a gene-by-environment interaction. In either location, different modifiers of the disease could be present (like temperature or chemicals) such that in one location the disease is or is not expressed when alleles on a certain haplotype are inherited.*

Question 9:
You are running a case-control GWAS for Type 2 Diabetes. Of the 500,000 variants you test, one variant (rs4514, which has 2 alleles, A and G) near the *sweetums* gene has good separation between cases and controls. You have 1000 cases, (480 of which are AA, 400 are AG, and 120 are GG at rs4514) and 1000 controls, (360 of which are AA, 440 are AG, and 200 are GG at rs4514).

(a) Using a chi-squared test, what is the p-value of the association of these alleles with the disease.

*A in cases = 480 * 2 + 400 = 1360, G in cases = 400 + 120 *2 = 640*
*A in controls = 360 * 2 + 440 = 1160, G in controls = 440 + 200 *2 = 840*

chi-squared statistic is 42.5 with a p-value of 7.17E-11

(b) Given that you did 500,000 tests, what is your (Bonferroni) corrected threshold for p-value significance (initial α=0.05)? Does the rs4514 variant pass "genome-wide significance" for association with Type 2 Diabetes?

0.05 / 500,000 = 1E-7

Yes, the p-value in part (a) is significant.

(c) What is the odds ratio of this variant in a risk for Type 2 Diabetes?

GRR_{AA} *= (AA cases / AA controls) / (GG cases / GG controls) = (480/360) / (120/200) = 2.22*

GRR_{AG} *= (AG cases / AG controls) / (GG cases / GG controls) = (400/440) / (120/200) = 1.51*