Name:______________________________________

Bio393

Biomedical Genetics

Final exam

Friday June 7, 2019

Graded exams will be available after 4 PM on Wednesday June 12th outside Pancoe 4115. If you have any questions about the grading of the exam, return your exam with a written explanation by Friday June 14th at noon. The grade distribution and key will be available on the course website.

Thank you for a fun quarter! Enjoy your summer break and/or your next adventure!

Please fill out the course CTECs. This year's assessment is especially important for Prof. Andersen.

Question 1 (8 points):

Genetic loci with large phenotypic effects are usually rare across populations.

(a, 4 points) Provide an explanation for why.

Genetic loci with large phenotypic effects are rare because in most cases they deleteriously affect fitness. Decreased fitness means that alleles are less likely to get passed on and do not increase in frequency across populations.

(b, 4 points) The variants that cause age-related macular degeneration have reached intermediate frequency in the human population. How do you think that large-effect variants (like those alleles) reach intermediate frequencies in humans?

The variants that cause AMD likely do not affect fitness so they can be passed on and can increase in frequency across populations. AMD is age-related, so most people will have offspring before these alleles can affect fitness as well. Additionally, AMD (even at its most severe case) will not cause lethality. Blind people can still have offspring, and the alleles will persist in populations.

Question 2 (6 points):
As an executive at a big pharmaceutical company, you want to sell drugs to as many people in as many countries as possible. Describe how you would design genome-wide association studies across multiple different populations (*i.e.* Europeans, Asians, Africans, etc.) to identify disease genes for a common genetic disorder.

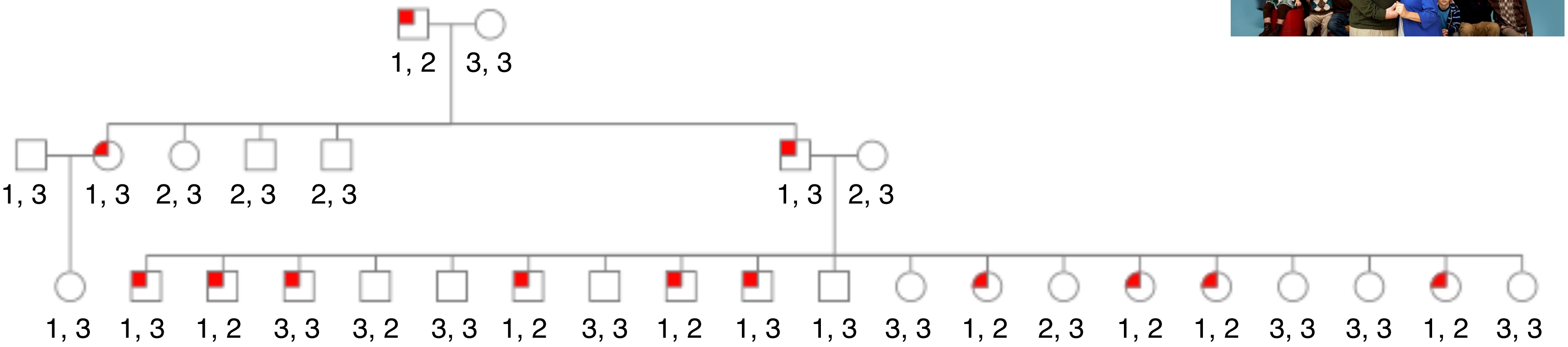
Because you would like to find a drug target that is common across many different populations, you should focus on GWA studies that find similar loci in all (or most) populations. However, you have to perform GWA studies in individual populations because otherwise population structure might confound your results.

You will perform GWA studies in each population separately and then compare the results. Any QTL that are shared across all independent GWA studies offer the best possible targets for future drugs that will work across diverse human populations.

Question 3 (24 points):

The TLC show *19 kids and counting* offers a rare opportunity to perform linkage mapping in a very large family. The following disorder (*gloriae cupiditate*) is observed in this family. Individuals marked with red have the disorder. You would like to test linkage between a marker (with three different alleles) and the disorder-causing allele (D). Choose a theta that will maximize the LOD score for this pedigree and fill in the equation for the LOD score for the entire family. The equation below is given as a guide.

$$\text{Odds ratio } = \log_{10} \frac{0.5*((1\text{-theta})^{P} * (\text{theta})^{R} + (1\text{-theta})^{R} * (\text{theta})^{P})}{(0.5)^{P+R}}$$



The gloridae cupiditate disease appears to be autosomal dominant making linkage mapping easier to assess.

The marker allele hypothesized to be nearby the disease-causing allele is 1.

To calculate the LOD score for the entire family, we need to divide the family up into phased and unphased sub-families and then add the LOD scores at the same theta.

Family #1 covers generations I and II. We don't know the phase of I-1, so we need to use the unphased LOD equation. Assuming phase 1 D, no recombinants are observed out of five children. So all offspring would be recombinants in the other phase.

Family #2 covers generations II and III and is on the left of the pedigree. Unfortunately, individual III-1 is uninformative so we need to exclude her from our calculations.

Family #3 is the big family that covers generations II and III. We know the phase of Dad II-6, so we can use the phased LOD equation. His phase is 1 D, so it looks like we have 2 recombinants and 17 parentals. We will use a theta of 2/19 for the entire family and both LOD equations.

Family#1

$$\text{LOD}_{\text{theta=2/19}} = \log_{10} \frac{1/2 * ((1\text{-}2/19)^{5} * (2/19)^{0} + (1\text{-}2/19)^{5} * (2/19)^{5})}{(1/2)^{5}}$$

Family#3

$$\text{LOD}_{\text{theta=2/19}} = \log_{10} \frac{(1\text{-}2/19)^{17} * (2/19)^{2}}{(1/2)^{19}}$$

$$\mathbf{LOD_{Total} = LOD_{Family\#1} + LOD_{Family\#3}}$$

Question 4 (14 points):
Crohn's disease affects about 0.001% of the European population. Given the increased prevalence of the disease in identical twins as opposed to fraternal twins, you suspect that this disease has a genetic cause. You perform a genome-wide association in a population of one million Europeans to identify genetic markers for Crohn's disease.

(a, 6 points) You performed the association mapping and found that the G allele at a SNV, *rs4077616*, is highly correlated with the disease. The ratio of the G to the other allele in the control population is approximately 1:1. Explain how individuals with the G allele in the control population may not be affected by the disease.

Individuals with the G allele in the unaffected population could not have the disease because the disease could be complex with many environmental and genetic effects. For example, an intergenic suppressor could be present in unaffected individuals, which reduces the likelihood of disease in those people. Also, the disease allele could have arisen in an individual with the G allele haplotype, but other individuals have the G allele without the disease-causing allele. The two are linked, but the G allele itself may not be causative. Finally, individuals in the control population with the G allele could have the disease-causing allele but low penetrance could make these individuals appear unaffected.

You decide to study 2000 individuals from your earlier association mapping that carry the G allele at *rs4077616* (1000 of which have Crohn's disease and 1000 of which do not have Crohn's disease). You identify a new SNV that is correlated with Crohn's disease in this smaller population. The A allele frequency is 40% and the T allele frequency is 60% in the population of affected individuals. The A allele frequency is 85% and the T allele frequency is 15% in the population of unaffected individuals.

(b, 8 points) Fill in the contingency table to set up a chi-squared test for the new SNV in this smaller population.

	cases	*controls*
A	*800*	*1700*
T	*1200*	*300*

Question 5 (12 points):
After taking BIO393, you decide to become an intern at Ancestry.com. Your first task is to calculate the genotype relative risk (GRR) for the CC genotype at a variant site implicated in high body mass index (BMI).

(a, 4 points) Using the table below, please calculate the GRR for the CC genotype in high BMI.

Genotype	Users with High BMI	Users with Normal BMI
CC	5000	2500
GC	20000	20000
GG	6000	8000

GRR_{CC} = (5000/2500) / (20000/20000) = 2

We normalized the GRR_{CC} by GRR_{GC} because it is the most abundant genotype in the dataset. You could normalize by GRR_{GG} to get a value of (5000/2500) / (6000/8000) = 2.67.

(b, 4 points) Please explain what the GRR_{CC} that you calculated in part (a) means in layman's terms.

This risk of having a high BMI is twice the average of the population if you have the CC genotype.

(c, 4 points) 23 and me also offers an assessment of the GRR_{CC} genotype at the same position in high BMI. They find that the GRR_{CC} value is 6.0. Please explain why the two companies might get different results for the same risk genotype.

The two companies have sampled different populations of individuals. The 23andme population could also have a different average genotype and risk ratio for that average genotype than the ancestry.com population.