Bio393: Genetic Analysis **Name:______________________**
Problem Set #7
Due on Friday, March 9, 3 PM

Question 1:
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 2:
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 3:
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

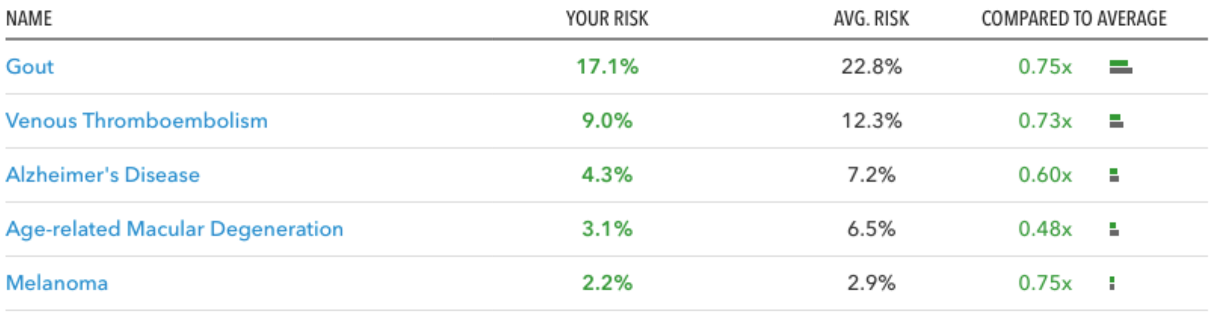
Yes, the p-value in part (a) is significant.

(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

Question 4:
A patient comes into your medical office presenting his 23andme results and an extreme sense of worry. He is completely confused about how 23andme determined that he has a reduced risk for gout, especially because he loves fatty foods. Please briefly explain in words how his risk can be less than 1x and how they calculated it.



NAME	YOUR RISK	AVG. RISK	COMPARED TO AVERAGE
Gout	17.1%	22.8%	0.75x
Venous Thromboembolism	9.0%	12.3%	0.73x
Alzheimer's Disease	4.3%	7.2%	0.60x
Age-related Macular Degeneration	3.1%	6.5%	0.48x
Melanoma	2.2%	2.9%	0.75x

23andme calculated his gout risk by looking at his genome-wide genotype. This individual shares haplotypes (markers) with individuals that have gout less often than the average person.

The alleles present in an individual might provide some protective benefit for certain diseases.