Molecular and Genetic Epidemiology
Assignment 5
Due by noon 2/10 by email to Kirsten Kangelaris, kkangelaris@medicine.ucsf.edu

A. Textbook, Read Austin Chapter 6

B. Project
We will continue to explore the potential genetic basis of "your" disease of interest. For this assignment, search PubMed for your disease and the terms "association study", "candidate gene", and "population stratification". (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi).

How many, if any, articles are on PubMed? If any, briefly glance at some of the abstracts and write a couple of sentences summarizing what has been done. Give a qualitative assessment of whether or not these findings provide a positive or negative support for a genetic basis to the disease.

C. Problems
1. A case/control study has been conducted and a SNP genotyped. Compute the odds ratios for the table below and the corresponding confidence intervals. Compute chi square tests for codominant, dominant and recessive modes of inheritance. Discuss the results in terms of plausibility of a model.

	X=0	X=1	X=2
Cases	500	350	120
Controls	521	270	130

Note: to get the OR and the CI in STATA, one can use:
. cci (# exposed cases) (# unexposed cases) (# exposed controls) (# unexposed controls)

Co-dominant model

	X_0	X_1	X_2	Total
Case	500	350	120	970
Control	521	270	130	921
	1021	620	250	1891
Odds Ratios	1	1.35	0.96	

Overall Expected

523.7	318.0	128.2
497.3	302.0	121.8

Chi-square p-value 7.1E-03

Dominant model

	X_0	X_1	X_2	Total
Case	500	470	0	970
Control	521	400	0	921
	1021	870	0	1891

Odds Ratios 1 1.22

Overall Expected

523.7	446.3	0.0
497.3	423.7	0.0

Chi-square p-value 2.8E-02

Recessive model

	X_0 & X_1	X_2		Total
Case	850	120	0	970
Control	791	130	0	921
	1641	250	0	1891

Odds Ratios 1 0.86

Overall Expected

841.8	128.2	0.0
799.2	121.8	0.0

Chi-square p-value 2.6E-01

Based on the test results and the trends in the OR it seems heterozygous disadvantage is a plausible model.

2. Analyses

Use the subset of the MTHFR SNP data that only includes the brothers discordant for prostate cancer from last week, but now test for association between the SNPs and prostate cancer ignoring the family relationships.

a. Open this file (on completion of downloading, or by double clicking on it), to launch STATA. Or start STATA and the command
. use "C:\MTHFRcase_control.dta", clear

b. Open a Log in STATA by going to **File, Log, Begin**. Give the log a name (suggest using the ".log" format).

c. Calculate the association between the SNPs and prostate cancer:
. tabulate cap mthfr_c677t_new, chi2 column
. tabulate cap mthfr_a1298c_new, chi2 column
Is there an overall association between the SNPs and prostate cancer? Comparing these results to those from last week, do you think the results are confounded by population stratification?

There is no overall association. The similarity in associations seen in both the unmatched and family-based analyses suggests that there is not bias due to population stratification.

3. Haplotypes

Recall that in the example dataset, the two MTHFR SNPs (C677T, A1298C) were strongly associated with each other, and the genotypes 677TT and 1298CC were never observed together. This suggests that we should consider these neighboring genotypes as haplotypes (i.e., set of alleles closely linked on one chromosome and inherited as a unit from a parent).

a. The following list gives joint genotypes for some of the individuals in the dataset. What are their (potential) haplotypes?

ID	MTHFR_C677T	MTHFR_A1298C	Haplotypes?
959	CC	AA	C-A / C-A
1044	CC	AC	**C-A / C-C**
147	CT	AA	**C-A / T-A**
123	CT	AC	**C-C / T-A or C-A / T-C**

b. When the two SNPs are heterozygotes (677CT and 1298AC), we don't necessarily know which SNP is on which chromosome, and so can't say what the haplotypes are. Any ideas about how we can figure out or estimate this?

We know that 677TT and 1298CC are never observed together, so it is unlikely that the haplotype T-C is observed. This suggests that the last haplotypes are C-C / T-A. One could further assign probabilities of the actual haplotypes by determining the haplotype frequencies from the known haplotypes. Information on the parent's genotypes would help as well.

c. Evolutionarily, what might explain the fact that the 677TT and 1298CC genotypes are never observed together?

This suggests that the original ("ancestral") haplotype at this location was C677 – A1298. Over time, new mutations may have been introduced, leading to C677 – 1298C and 677T – A1298 haplotypes. Due to the rarity of these mutations, the fourth possible haplotype, 677T – 1298C was never observed. This haplotype would not be expected if the haplotype was associated with a strong survival disadvantage. Also, if the two mutations occurred separately there may not have been enough time since they occurred for them to be inherited on the same chromosome.