Problem Set 2
BENG 183
Fall 2011 Due Tuesday October 11th 2011, 11:59 PM
Deliver in class Tuesday or submit by e-mail to q1ma@ucsd.edu.

1) In the post-genomic era, there has been a substantial push for personalized medicine. With ever decreasing costs in genome hybridization and sequencing techniques, the process is now affordable for the average consumer. 23andMe is one of many companies offering these services.

   a. Go to the 23andMe website (http://www.23andme.com) and register a demo account. First, click on "How it Works" on the homepage and then "try a demo" on the successive page. Make sure to email Qi Ma the email address used for your registration. It is needed for the 23andMe staff to unlock certain educational features.

   b. In order to get you familiar with the interface and navigation of the 23andMe website, answer the following questions:
      i. Which disease is Greg Mendel most at risk for? Which disease does he have the highest risk compared to the average population? Which disease does he have the lowest risk compared to the average population?
      ii. In a genome-wide comparison, who is Alan Mendel's SNPs most similar to in the family? Who is he most similar to outside the family?
      iii. Report the percent similarity of Lilly Mendel with Margo Fisher and Erin Mendel for Circadian Rhythm and Pigmentation related SNPs? Can you think of a reason(s) of the discrepancy?
      iv. 23andMe is improving its services with new features. What are two research highlights that are in the pipeline?

2) Principal component analysis (PCA) is a mathematical tool to explain the variance in a data set. PCA transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The first calculated principal component accounts for the greatest amount of variability in the data set. Successive components account for the remaining variability. We will use the method to find the most correlated SNPs in defining a particular group (in this case 12 patients). Read Paschou et al. PLoS Genetics 2007 (PMID: 17892327) for more details.

   a. You will need to use Matlab for this problem. Download the data file and readData.m from the Assignments page on the website. Use readData.m to load the combined patient SNPs. The generated matrix (**A**) is **m** x **n** with rows of **m** representing subjects and columns of **n** representing the different SNPs. A heterozygous genotype is represented as a 0, while 1 and -1 are assigned to the two potential homozygous genotypes.

b. We will now calculate the SNPs that best represent the population (12 patients). We will use singular value decomposition (svd in Matlab) to do so. In brief, a matrix (**B**) is decomposed into three new matrices:

$$B = U\Sigma V^T$$

In order for the svd calculation to be computationally tractable, calculate the svd of the covariance matrix: **A*****A**^{T}. Using the equation above, analytically show that the singular value decomposition of the covariance matrix is:

$$A \cdot A^T = U\Sigma^2 U^T$$

c. The correlation scores (**p**) for each SNP can be calculated using the following formula:

$$p_j = \sum_{i=1}^{m} (v_j^i)^2$$

where each v_j represents a column of V. Using the calculated properties above, determine the matrix **V** and **p** for each SNP.

d. To turn in, report the top 5 SNPs and the values of their scores. Also, show the property in (b) and print out and attach any code used for the calculation.

3.) Computing a Log Odd (LOD) score can be useful for determining linkage of a SNP to a particular phenotype. In the progeny below, there are 10 individuals' SNPs that are thought to be correlated with a particular disease gene (Diseased/D and Normal/N).

Individual	1	2	3	4	5	6	7	8	9	10
SNP	'AC'	'AC'	'AC'	'AC'	'AA'	'AA'	'AA'	'AC'	'AA'	'AC'
Phenotype	D	N	N	N	D	D	N	N	D	N

a. What seems to be the SNP of the diseased and normal phenotypes?

b. What is the probability of the observed phenotypes that the SNP is completely unrelated to the disease gene?

c. What is the probability of the observed phenotype that there is a 10% chance of crossover of the SNP as compared to disease gene?

d. Compute the LOD score?

e. Compute the LOD score for 20 and 30% recombination?

f. From the calculated LOD scores, what is the optimal recombination fraction, maximizing the LOD score?

g. What is the probability of having the disease phenotype with the genotype ‘AC’, with the genotype ‘AA’. What is the odds ratio for getting the disease with genotype ‘AA’ versus ‘AC’?