Genetics 210
Problem Set 1
Due: April 19, 2012 to gene210.stanford@gmail.com

NOTE: Each question was worth 6.25 points

1. Single nucleotide polymorphisms (SNPs) are a significant source of human genetic variation. Currently, the most common approach to determine SNPs present in an individual's genome is through high-density SNP microarrays. These "SNP chips" are popular due to their high-accuracy and relatively low costs, and as such, several direct-to-consumer genotyping companies employ this approach to provide SNP genotypes at hundreds of thousands of variable sites within the human genome. As will be demonstrated throughout this course, many of these sites provide information regarding an individual's disease susceptibility and drug response, in addition to information about ancestry and heritable traits. The following is a primer to familiarize you with this type of genetic data:

A. Go to **dbSNP** (http://www.ncbi.nlm.nih.gov/projects/SNP/) and find the "dbSNP Summary" under the "General" section on the left panel.

1. When was the current human build (135) last updated?
   Late 2011

2. How many human SNPs have been identified *and validated* as of build 135? (Found in "Build Statistics" in the "Number of RefSNP Clusters" column)
   ~ 41 million

3. How many unique human SNPs have been identified and validated *since* the build 135 update? (Found in "New Submission since previous build" in the "New RefSNP Clusters" column)
   ~ 13 million

B. Compare the number of identified and validated SNPs from question "A.2" to the *total* human genome size[1]. At what percent of the genome have polymorphisms been observed?

(41,740,143 / 3,137,161,264) * 100 = **1.33050677%**

C. If you were to re-sequence a new human genome (say, Steve Quake's[2]), approximately how SNPs would there be relative to (percentage):

1. The size of the human genome

[1] http://genomewiki.ucsc.edu/index.php/Hg19_Genome_size_statistics
[2] Ashley, E.A. *et al*. Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525-35 (2010).

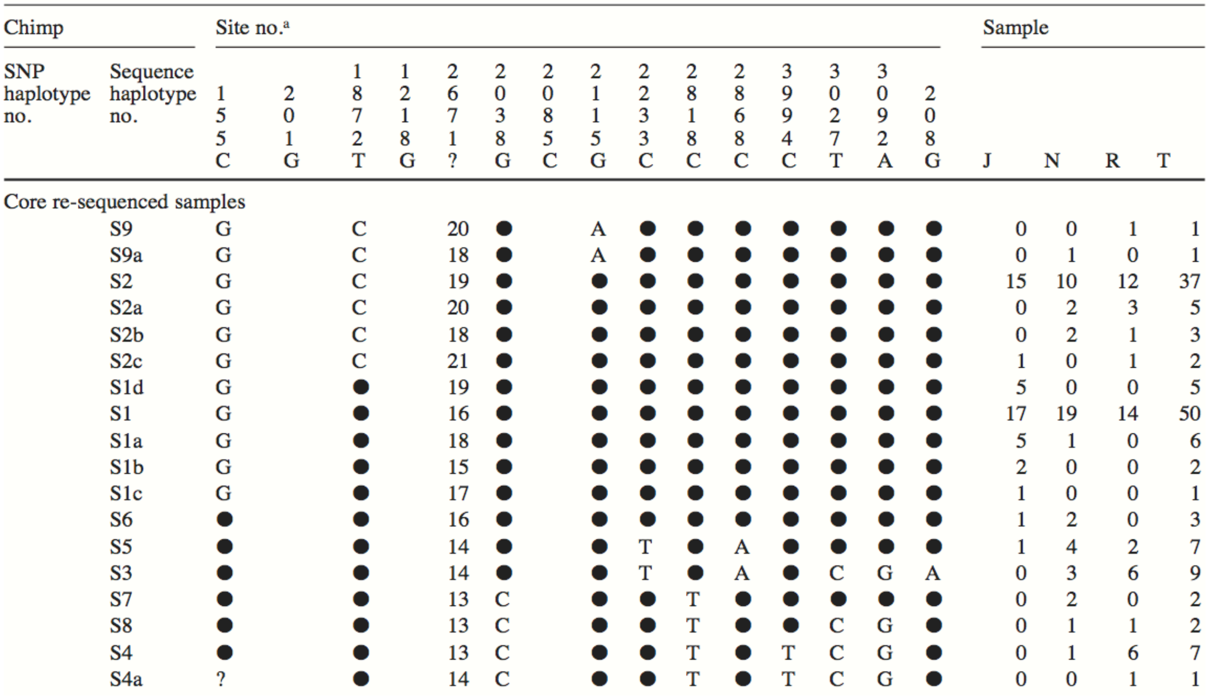
(2,800,000 / 3,137,161,264) * 100 = **0.0892526639%**

2. The total number of observed and verified SNPs (from part “A.2”)?

   (2,800,000 / 41,740,143) * 100 = **6.7081706%**

2. The following table shows the genotypes of APOA2 from 72 people. The last column (T) is the total number of chromosomes that contain that haplotype. The sequence at the top is the reference sequence. A circle indicates match to reference. Position 2671 is a simple copy number repeat, and can be ignored in this example. **All haplotypes that differ only at 2671 should be considered as one.**

You read an interesting paper about a SNP at position 3092 in the APOA2 gene. However, your DNA chip only contains a SNP at position 208. You want to know how well you can impute your genotype at position 3092 using your genotype at position 208. To do this, you need to evaluate whether these two alleles show linkage disequilibrium. Information on linkage disequilibrium is in the class notes and at http://en.wikipedia.org/wiki/Linkage_disequilibrium.



Chimp		Site no.[a]															Sample			
SNP haplotype no.	Sequence haplotype no.	1555 C	201 G	1872 T	1218 G	2671 ?	2038 G	2085 C	2115 G	2233 C	2818 C	2868 C	3994 C	3027 T	3092 A	208 G	J	N	R	T
Core re-sequenced samples																				
	S9	G		C		20	●		A	●	●	●	●	●	●	●	0	0	1	1
	S9a	G		C		18	●		A	●	●	●	●	●	●	●	0	1	0	1
	S2	G		C		19	●		●	●	●	●	●	●	●	●	15	10	12	37
	S2a	G		C		20	●		●	●	●	●	●	●	●	●	0	2	3	5
	S2b	G		C		18	●		●	●	●	●	●	●	●	●	0	2	1	3
	S2c	G		C		21	●		●	●	●	●	●	●	●	●	1	0	1	2
	S1d	G		●		19	●		●	●	●	●	●	●	●	●	5	0	0	5
	S1	G		●		16	●		●	●	●	●	●	●	●	●	17	19	14	50
	S1a	G		●		18	●		●	●	●	●	●	●	●	●	5	1	0	6
	S1b	G		●		15	●		●	●	●	●	●	●	●	●	2	0	0	2
	S1c	G		●		17	●		●	●	●	●	●	●	●	●	1	0	0	1
	S6	●		●		16	●		●	●	●	●	●	●	●	●	1	2	0	3
	S5	●		●		14	●		●	T	●	A	●	●	●	●	1	4	2	7
	S3	●		●		14	●		●	T	●	A	●	C	G	A	0	3	6	9
	S7	●		●		13	C		●	●	T	●	●	●	●	●	0	2	0	2
	S8	●		●		13	C		●	●	T	●	●	C	G	●	0	1	1	2
	S4	●		●		13	C		●	●	T	●	T	C	G	●	0	1	6	7
	S4a	?		●		14	C		●	●	T	●	T	C	G	●	0	0	1	1

A. Given the above data[1], calculate the allele frequencies for position 3092 and position 208.

3092: f(A) = 125 / 144 = 0.868, f(G) = 19 / 144 = 0.132.

208: f(G) = 135 / 144 = 0.938, f(A) = 9 / 144 = 0.0625.

[1] Fullerton, S.M. et al. Sequence polymorphism at the human apolipoprotein AII gene (APOA2): unexpected deficit of variation in an African-American sample. *Hum Genet* **111**, 75-87 (2002).

B. Next, calculate the haplotype frequencies from alleles at position 3092 and position 208.

Note that answers are in (3092, 208) order.

f(AG) = 125 / 144 = 0.868.

f(GG) = (2 + 7 + 1) / 144 = 0.0694. From the S8, S4 and S4a haplotypes.

f(GA) = 9 / 144 = 0.0625. From the S3 haplotype.

C. Calculate D‘ between position 3092 and position 208.

To start, calculate the disequilibrium coefficient between 3092 and 208, D(3092, 208). This is f(AG) – f(A) f(G).

From above, f(AG) = f(A) = 0.868. f(G) = (144 – 9) / 144 = 0.938.Then, D = 0.868 – 0.868 * 0.938 = 0.0543.

Since D’ = D / min(p(Ab), p(aB)), we need to calculate the denominator. This is min(0.868 * 0.0625, 0.132 * 0.938) = 0.0543. Thus D’ = 0.0543 / 0.0543 = 1, indicating that we never observed the AA haplotype.

D. Calculate R^2 between position 3092 and position 208.

D ^ 2 = D(3092, 208) ^ 2 / f(A) (1 - f(A)) f(B) (1 - f(B)). We calculated D(3092, 208) above. Then we have 0.0543 ^ 2 / (0.868 * 0.132 * 0.938 * 0.0625) = 0.439.

E. Based on the number of sites in the table**,** how many haplotypes are possible? Ignore positions 155, 201, 1218, 2671 and 2085 (where the data are missing or incomplete) in this problem. Assume that the polymorphisms segregate randomly with respect to each other.

In this problem, we wanted you to simply look at the number of sites (15 minus the 5 we are excluding). Since there are two base possibilities at each of those sites, we have a total of 2 ^ 10 = 1024 possible haplotypes.

F. How many haplotypes are observed in the set of 144 sequenced chromosomes from the table? What is the reason for the difference between the observed number of haplotypes and the total possible number?

Both 8 and 9 were acceptable answers here (8 if you observed that S1 and S6 are identical the sites we instructed you to include). This difference is due to linkage disequilibrium – stretches of DNA tend to travel together unless interrupted by recombination, which occurs less frequently at smaller between-site distances.

G. The third row shows the S2 haplotype. What is the expected haplotype frequency for S2 if all of the SNPs segregated randomly with each other? What is the observed frequency for the S2 haplotype?

For this question, we expected you to calculate frequency as below. However, some took this to essentially be a repeat of question E, in which cases they calculated the expected frequency as 1 / (2 ^ n), where n is the number of sites. Partial credit was given for a response of this form, with n either 10 (if you continued to ignore some of the sites) or 15 (if you did not). In this case, the answers were 0.000977 (for 10 sites) or 3.05E-5 (for 15 sites).

The method we were looking for was to find the expected haplotype frequency from per-allele frequency. For this calculation, at each site, you need to count the total frequency of the S2 haplotype's allele in the entire sample (of 144 haplotypes).

Then, your terms for each of the ten sites are 49, (144 - 12), (144 - 2), (144 - 16), (144 - 12), (144 - 16), (144 - 8), (144 - 19), (144 - 19) and (144 - 9). To get allele frequencies, each of these is divided by 144.Finally: 144 ^ -10 * 49 * 132 * 142 * 128 * 132 * 128 * 136 * 125 * 125 * 135 = 0.149, which is your expected frequency for the S2 haplotype.

Some people did not ignore site 155 (which had a missing value in the last row). In this case, the expected frequency would have been 0.149 * 113 / 144 = 0.117.

The observed frequency is simply the sum of all S2 haplotypes – that is, (37 + 5 + 3 + 2) / 144 = 47 / 144 = 0.326.

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 *SUGAH* 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).

A. Using a chi-squared test, what is the p-value of this association?

We weren't entirely clear in this subquestion whether we wanted you to give p- values for genotypes or alleles. Stuart also demonstrated an alternative model for this hypothesis in class, using the control group for expected counts. Any of these four possible approaches was acceptable. First for genotypes and alleles by the more standard model, using R:

```
# Test genotypes. genotypes <--- matrix(c(480, 400, 120, 360,
440, 200), byrow = TRUE, nrow = 2)
```

```
print(chisq.test(genotypes))
```

The chi-squared statistic is 39.0476, yielding a p-value of 3.32E-9 at two degrees of freedom.

As in the cases: 480 * 2 + 400 = 1360

Gs in the cases: 400 + 120 * 2 = 640

As in the controls: 360 * 2 + 440 = 1160

Gs in the controls: 440 + 200 * 2 = 840

```
# Test alleles.
```

```
alleles <--- matrix(c(1360, 640, 1160, 840), byrow = TRUE, nrow
= 2)
```

```
print(chisq.test(alleles)) # correct = TRUE is the default.
```

The chi-squared statistic is 42.5 (or 42.9 if you specified correct = FALSE), yielding a p-value of 7.17E-11 (or 5.76E-11) at one degree of freedom.

And with the controls-as-expected models:

```
# Test genotypes.
```

```
cases <--- c(480, 400, 120)

controls <--- c(360, 440, 200)

print(sum((cases --- controls) ^ 2 / controls))
```

The chi-squared statistic is 75.6, with two degrees of freedom.

```
# Test alleles.

cases <--- c(1360, 640)

controls <--- c(1160, 840)

print(sum((cases --- controls) ^ 2 / controls))
```

The chi-squared statistic is 82.1, with one degree of freedom.

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?

If your desired alpha (false discovery rate) is the same as your p-value, you just adjust your p-values by the number of hypotheses you are testing. In other words, any variant must have a p-value less than 0.05 / 5E5 = 1E-7. Regardless of which way you chose to approach testing, the p-value for rs4514 is low enough to be significant even after multiple hypothesis correction.

C. What is the odds ratio of this variant in a risk for Type 2 Diabetes?

Recall that the odds ratio is of the form odds(A) / odds(G).Odds(A) = number of As in the cases / number of As in the controls = 1360 / 1160 = 1.17

Odds(G) = number of Gs in the cases / number of Gs in the controls = 640 / 840 = 0.762.

Finally, the odds ratio is 1.17 / 0.762 = 1.54.