SISG 2022: Module 11
Session 4: Hardy-Weinberg Equilibrium and Linkage Disequilibrium

1. Sickle Cell Anemia is characterized by fatigue, pain, arthritis, frequent bacterial infections, and sudden pooling of blood in the internal organs that can lead to tissue damage. It is caused by a SNP in the hemoglobin gene that causes red blood cells to form sickle shapes instead of round, donut shapes. The SNP is a missense mutation (T>A) that replaces a glutamine with a valine and causes hemoglobin molecules to clump.

   a) You are conducting a study and find that the A allele has a frequency of 20% among adults. Use Hardy-Weinberg Equilibrium equations to calculate how many people out of 1,000 you would expect to have each genotype. Remember the HWE equations are

   $$1 = p + q$$
   $$1 = p^2 + 2pq + q^2$$

   b) In the study of these same 1,000 people, you find that actually 605 people have the TT genotype, 390 have the TA genotype, and 5 people have the AA genotype. Is this a statistically significant deviation from what you would expect based on the allele frequencies? We calculate the chi-square value with the following equation and compare it to a chi-square distribution with 1 degree of freedom (3.841).

   $$x^2 = \Sigma \frac{(O_i - E_i)^2}{E_i}$$

   c) What might be happening in the population to give you this HWE pattern?

2. Why do we only look for LD between SNPs that are on the same chromosome?

3. Here we will explore LD using the NCI LDLink online tools. You can find this website at ldlink.nci.nih.gov/?tab=home. We have a lot of different tools to explore, but here we will use the LDpop tab.
   a. Compare LD for the two SNPs that define alleles in two important genes affecting drug metabolism. These are rs776746 and rs2740574. Type these into the two SNP boxes (variant RS number) and select "(ALL) all populations", "R^2", then "calculate." After a few seconds, you should see a map of the world with tear drops showing different populations that have been studied, each labeled by the population. What is the LD R^2 value among the British in England and Scotland (GBR) compared to the LD R^2 value among the Luhya in Webue, Kenya (LWK) and compared to among Colombians from Medellin, Colombia (CLM)? You can find the details for each population in the table, and by clicking on the corresponding tear drop.

   b. Why might these LD values be so different between these populations?

   c. Most genetic studies occur in participants of European descent. When we are studying drug metabolism affected by these two SNPs, why might knowledge from these studies have limited utility across ancestral populations?

1) Below is a plot of principal component 1 vs principal component 2 in a sample of people from 4 populations: African Americans from the Southeast (AAS), Europeans from Utah (CEU), Yorubans from Nigeria (YRI), and Han Chinese from Beijing (HCN). Each dot represents one person and each person is color-coded based on their self-described race.

PCA in Reference Sample
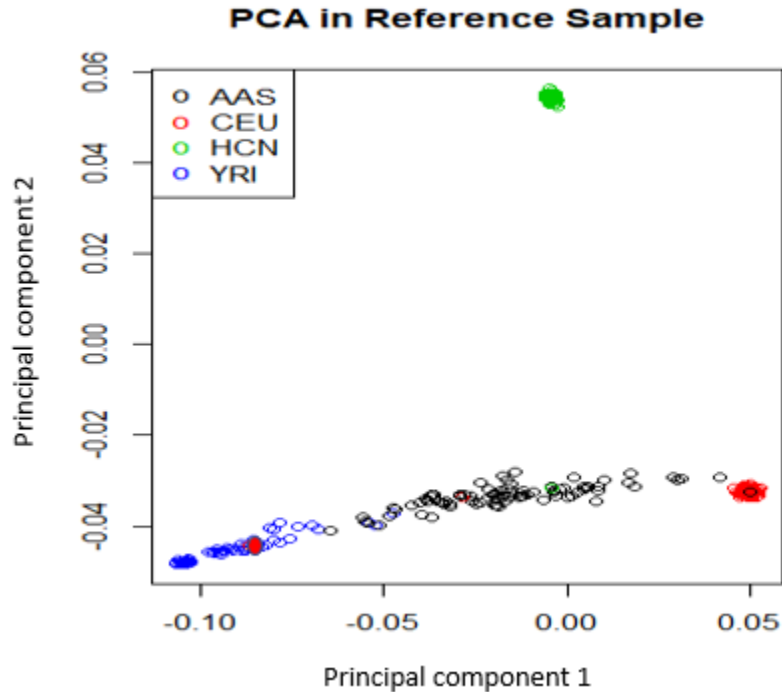
a) What populations are separated by principal component 1? What populations by principal component 2?

b) Why do we see tight-ish clusters of the three corner populations (blue, red, green), but the black circles are spread across the axis from blue to red along principal component 1?

c) Notice the red dot in the lower left corner among the blue dots. What might be happening here? Remember what color refers to compared to what the principal components measure.

d) Where on this plot might you see people who describe their ancestry as Chinese American (ancestors from both European and Chinese populations)?

e) What are pros/cons of using self-described race vs genetic ancestry in epidemiology studies? Think of what each can tell you based on the questions you are trying to ask.

SISG 2022: Module 11
Session 6: Study design for genetic association studies

1) Explore the breakdown of genetic ancestry in GWAS as reported on the website https://gwasdiversitymonitor.com. What do you notice about recent trends? What populations seem over- and under-represented in genetic studies?

2) What are your ideas for how we can we increase the diversity of study participants in genetic epidemiology?

SISG 2022: Module 11
Session 7: Association studies calculations and interpretations

1. You conduct a case/control study among 1,656 participants. You are particularly interested in the odds ratio for the outcome among those homozygous for the C allele vs. either heterozygous or homozygous for the T allele. You genotype everyone for that particular SNP and find the following genotype frequencies among your cases and controls.

   a. Calculate the odds ratio and 95% confidence interval for the odds of having the outcome among CC vs TT/TC genotypes.

	Cases	Controls	Total
TT+TC	158	392	550
CC	20	86	106
Total	178	478	1,656

   Remember the following equations:

   $$OR = \frac{ad}{bc}$$

   $$s.e(\log(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

   Lower limit of 95% confidence interval: $e^{\log(OR) - 1.96 \times s.e}$

   Upper limit of 95% confidence interval: $e^{\log(OR) + 1.96 \times s.e}$

   b. Turn this result into a sentence describing the association between the CC genotype and odds of the outcome compared to the TT/TC genotypes.

2. You conduct a case/control study using an additive inheritance model. Your logistic regression output is as follows:
   Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.75	1.493	34.480	<2e-16***
genotypeAdd	1.504	0.251	6.714	<2e-16***

   a. Determine the odds ratio for the odds of the outcome among participants with 2 copies of the allele of interest (genotypeAdd =2) compared to the odds among participants with 1 copy of the allele of interest (genotypeAdd = 1).

   b. Use the std.error to determine the 95% confidence interval for that odds ratio estimate using the following equation with the standard error (s.e.):

c. Bonus: Determine the odds ratio for the odds of the outcome among participants with 2 copies of the allele of interest (genotypeAdd =2) compared to the odds among participants with no copies of the allele of interest (genotypeAdd = 0).

3. You conduct a quantitative association study of bone mineral density using an additive genotype model. Your linear regression output is as follows:

Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.75	1.493	34.480	<2e-16***
genotypeAdd	1.504	0.251	6.714	<2e-16***

a. What is the average change in bone mineral density for every additional allele of interest?

b. Among people homozygous for the allele of interest (genotypeAdd=2), what is the average bone mineral density?

SISG 2022: Module 11
Session 8: Genome wide association studies

1) Explore the NHGRI-EBI GWAS catalog: https://www.ebi.ac.uk/gwas/home. This website will introduce you to existing GWAS on many different phenotypes

2) Using the GWAS catalog, determine what the SNP rs6025 has been associated with in previous studies.

3) Explore the Global Biobank Engine (https://biobankengine.stanford.edu), which has collated GWAS results on a wide range of phenotypes based on large biobanks (UK Biobank, Biobank Japan, Million Veterans Program). Using this resource, what associations do you see with rs6025?