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

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

Assuming 20%A, 80% T:
Based on HWE: AA: $0.2^2 = 0.04$; TT = $0.8^2 = 0.64$; AT = 2x0.2x0.8 = 0.32
Estimated Genotype frequency: AA: 40; TT: 640; AT: 320.

   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).

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

Estimated frequency: AA: 40; AT: 320; TT: 640
Observed frequency: AA: 5; AT: 390; TT: 605
X2 = $[(5-40)^2/40 + (390-320)^2/320 + (605-640)^2/640] = 47.8516$
It is a statistically significant deviation from the HWE

   c) What might be happening in the population to give you this HWE pattern?
Sickle Cell Anemia is common in the same global locations where malaria has historically been very common (Africa, India, Middle East). The malaria parasite cannot survive in red blood cells that sickle. In fact, in these regions, children with the TA genotype are most likely to survive to adulthood. This is called the "heterozygote advantage." Children with the TT genotype are more likely to die from malaria, children with the AA genotype suffer from sickle cell anemia, and children with the TA genotype have natural defenses against malaria because their cells sickle under pressure (infection) but remain round when not stressed.

2. Why do we only look for LD between SNPs that are on the same chromosome?
Chromosomes are segregated independently during meiosis, so SNPs on different chromosomes are not able to be physically linked.

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.

R2 among GBR: 0.5864
LWK: 0.0276
CLM: 0.225

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

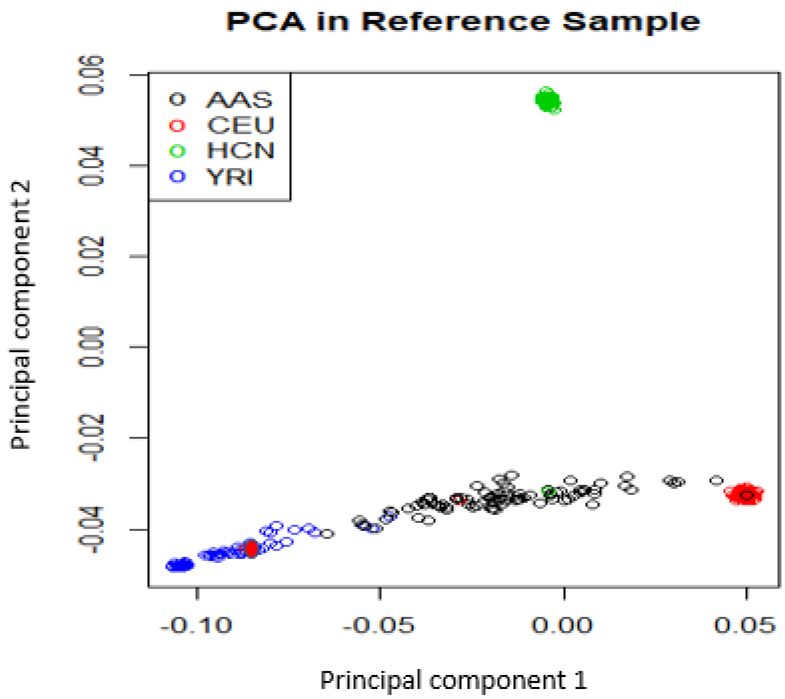
Because LD pattern can reflect the evolutionary history of a population and may be affected by various factors, including nature selection, genetic drifting, inbreeding, non-random mating, bottleneck effect, and founders' effect.

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?

The research results based on European-ancestry population may not be applicable to other populations, as the linkage between those two SNPs in other populations of interest is medium to low.

SISG 2022: Module 11
Session 5: Population substructure

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 group.



a) What populations are separated by principal component 1? What populations by principal component 2?

Most individuals have a negative PC2 value, but HCN population has positive PC2; AAS, CEU and YRI populations can be separated using PC1, although there are some potential misclassifications.

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?

AAS population is internally heterogeneous. Probably they are nearly admixed population that has features between YRI and CEU.

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.

Potential misclassification of self-described race/Distinctive results of self-described and genetically determined race.

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

Somewhere between the red and green cluster.

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.

Self-described race, pros: Easily obtained via interview/questionnaire; may reflect the character of race/ethnicity beyond genetic background (culture, shared lifestyle); cons: Potential misclassification as self-determined race may be distorted.

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

> Explore the breakdown of genetic ancestry in GWAS as reported on the website https://gwasdiversitymonitor.com.

  - What populations seem over- and under-represented in genetic studies?
  - What consequences can this have?

  Majority of the genetic association studies were based on European ancestry population. In recent years, although number of studies in Asian and African American population increased, the minority populations (Africans, African Americans, Asians, and Hispanics) are still under-represented. This can have many downstream consequences, as many therapeutics are developed based on genetic information. These treatments may be ineffective or even detrimental for those who are underrepresented in genetic studies. Overall, we also understand less about the biology of disease by limiting our studies of genetic variation.

> 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:

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

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

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

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

Odds ratio = (158x86)/(392x20) = 1.73.
Log(OR) = 0.5481
SE(log(OR)) = (1/158 + 1/392 + 1/20 + 1/86)^1/2 = 0.2655
Upper limit of the CI: exp(0.5481+ 1.96 x 0.2655) = 2.91
Lower limit of the CI: exp(0.5481- 1.96 x 0.2655) = 1.03

    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.

Relative to the population with CC genotype at the locus, those who with TT or TC genotype would have 1.73-fold of the odds to develop the outcome.

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).

Odds ratio = exp(1.504) = 4.50

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.):

Lower limit = exp(1.504 - 1.96 x 0.251) = 2.75
Upper limit = exp(1.504 + 1.96 x 0.251) = 7.36

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).

OR = exp(2 x 1.504) = 20.25

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?

REMEMBER WE ARE USING A LINEAR REGRESSION HERE.
For every additional allele of interest, the bone mineral density increases by 1.504 units

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

Average bone mineral density among rare homozygous = 7.75 + 1.504 x 2 = 10.758

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.
   Five trait(s) have been associated with SNP rs6025 based on the previous GWASs: venous thromboembolism; Antithrombotic agent use measurement; abnormal thrombosis, deep vein thrombosis, Ischemic stroke, pulmonary embolism, stroke, venous thromboembolism; inflammatory bowel disease; peripheral arterial disease

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?