CSCI 2820 HW1: Linkage Disequilibrium

Due date: 19 February 2021 Friday

Submission Instructions: Save your write-up to a PDF document titled ***"YourLastName-CS2820-HW1.pdf"*** and e-mail it to your graduate TA at pinar_demetci@brown.edu along with the supplementary material asked for (e.g. code), if any, by the deadline indicated above. Supplementary materials should also be similarly titled (e.g. "YourLastName-CS2820-HW1-Code.py") or included in your write-up.
All deadlines are ***11:59pm Anywhere on Earth***. While this corresponds to 7am the next day in EST, we do not want you to stay up late to turn in assignments. This choice of time is to accommodate anyone currently taking classes outside the U.S.

Notes: You have a ***total of 5 free late days*** you can use however you want on written and programming assignments this semester (except for the final project) without incurring any penalties. After that, you are penalized 10% for each late day.
If there is anything unclear, please ask questions on Piazza and come to our office hours!

1 Linkage Disequilibrium (LD)

1.1 Measures of LD

The table below gives the number of occurences for each haplotype. Calculate LD between locus A,a and C,c using each of D, D', and r^2. Please show your work.
Based on the r^2 value, is LD between these two loci statistically significant?

ABC	198	aBC	97
ABc	29	aBc	5
AbC	16	abC	6
Abc	11	abc	3

Do the same for the locus G,g and locus H,h below:

	H	h
G	3	16
g	14	5

1.2 Cases for LD

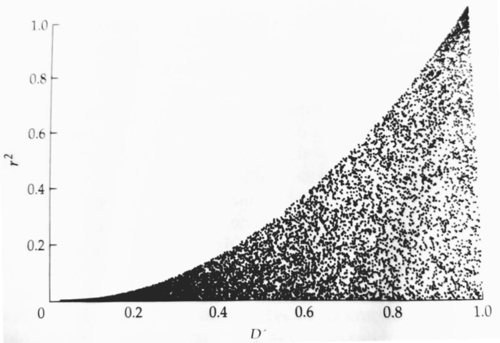
As the figure below captures (and as you can observe from the defitions of each measure and in-class examples), D' and r^2 do not always give the same values. For example, loci that are in perfect LD ($r^2 = 1$) are necessarily in complete LD ($D' = 1$) but not vice versa. If the two loci have very rare alleles and the rare alleles do not occur together on a haplotype, we might have a case of $D' = 1$ (since one of the possible haplotypes will not be observed in the samples) but a small r^2.

Given that, it's important to have some intuition about what your metrics measure and how they behave in different scenarios. To that end, construct example data where:

1. $r^2 = 1$ and $D' = 1$

2. $r^2 < 1$ and $D' = 1$
3. $r^2 < 1$ and $D' < 1$

Figure 1: Relationship between D' and r^2 for 10,000 random, uniformly distributed values of the gametic frequencies. Image taken from "Principles of Population Genetics" by Hartl & Clark, 4th edition, page 84.



2 Questions on assigned reading

Please skim the following paper (posted on course website among assignments):

- "Patterns of Linkage Disequilibrium in the Human Genome" by Ardlie et al.

and answer the following questions:

1. What is the leading cause for erosion of linkage disequilibrium? Please explain.
2. What are "useful LD", "complete LD", "perfect LD" and "LD blocks"? Explain the difference between complete LD and perfect LD in your own words. What is studying LD blocks useful for?

2.1 Seminal paper of the chapter

Please carefully read the following paper:

- "Linkage disequilibrium — understanding the evolutionary past and mapping the medical future" by Slatkin.

Explain your main takeaways from the reading. Some guiding questions are:

1. Why study LD? What can LD blocks or patterns of LD (throughout the genome, at specific loci, between populations, and within a population) can tell us? What are some current applications of LD?
2. (In part, related to the question above) What are the genetic, demographic, and selection-based factors that affect LD? Please explain.
3. What are some challenges and limitations of current LD measures (e.g. when studying history of population genetics, when studying more than two loci or multi-allelic loci, when carrying out LD mapping for GWAS)? According to the paper, what are some practices used to get around these? What are the advantages/disadvantages? Do you find them practical?
4. Are there any points of confusion after reading this paper?

3 Fisher's exact test

3.1 The story behind the test

Fisher's exact test is a statistical test used for the analysis of contingency tables. An exact test is a test where you can compute p values exactly (rather than approximating with asymptotic distributions, for example the Poisson distribution approximation for the binomial).

This test was developed by Fisher (hence the name) for a randomized experiment to test if a lady, who claimed she could determine whether or not milk or tea was added first to a cup, could indeed do that. This experiment is known as the *Lady Tasting Tea Test.*

Fisher gave the lady 8 randomly ordered cups of tea. Half of the cups were prepared first with milk (then tea) and half were prepared first with tea (and then milk). The null hypothesis is that the lady has no special ability to determine which cups are which.

The lady chooses 4 out of the 8 cups prepared by one method. The test statistic is the number of choices she got correct, $X \in \{0, 1, 2, 3, 4\}$. The reason that this is an exact test is because we can count the number of permutations (orderings) of the number of cups she got correct. This is summarized in the table

X	Permutations
0	1
1	16
2	36
3	16
4	1
Total	70

In Fisher's experiment the lady got all 4 cups correct. The critical region for this test (at significance $\alpha = 0.05$) is $X = 4$ since

$$P(X = 4) = \frac{1}{70} \approx 0.014$$

while

$$P(X \geq 3) = \frac{16 + 1}{70} \approx 0.243$$

which is not significant. So we reject the null hypothesis that the lady has no ability to choose between preparation methods.

3.2 Written Problem

1. Derive the probability of obtaining any particular set of values w, x, y, z in the 2×2 contingency table of alleles below for Locus 1 with alleles A, a and Locus 2 with alleles B, b

	B	b	Row Total
A	w	x	$w + x$
a	y	z	$y + z$
Column Total	$w + y$	$x + z$	$w + x + y + z = n$

2. Why doesn't this test scale to large sample sizes (i.e. why do we usually use the χ^2-test for contingency tables)?

3.3 Implementation

Implement Fisher's exact test in the programming language of your choice and apply it to the contingency table found below to test whether the genomic locus A/a and the locus B/b are in linkage disequilibrium. What is the null hypothesis, H_0? What is the associated p-value? Do we reject the null hypothesis?

	B	b	Row Total
A	12	16	28
a	8	5	13
Column Total	20	21	41

3.4 Comparison

Apply an appropriate χ^2 goodness of fit test to the above data and compare the p value you obtain with the p value obtained from Fisher's exact test.

Bonus (10 points)

Attempt to mathematically derive the relationship between χ^2 and r^2. Full and formal mathematical proofs will get full credit (10 points bonus) but partial credit will also be rewarded for incomplete derivations. **Tip:** Think of the definition of χ^2 in terms of expected and observed outcomes.