Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

Optional Lecture1: Concepts in Population Genetics and Haplotype Testing in GWAS

Jason Mezey

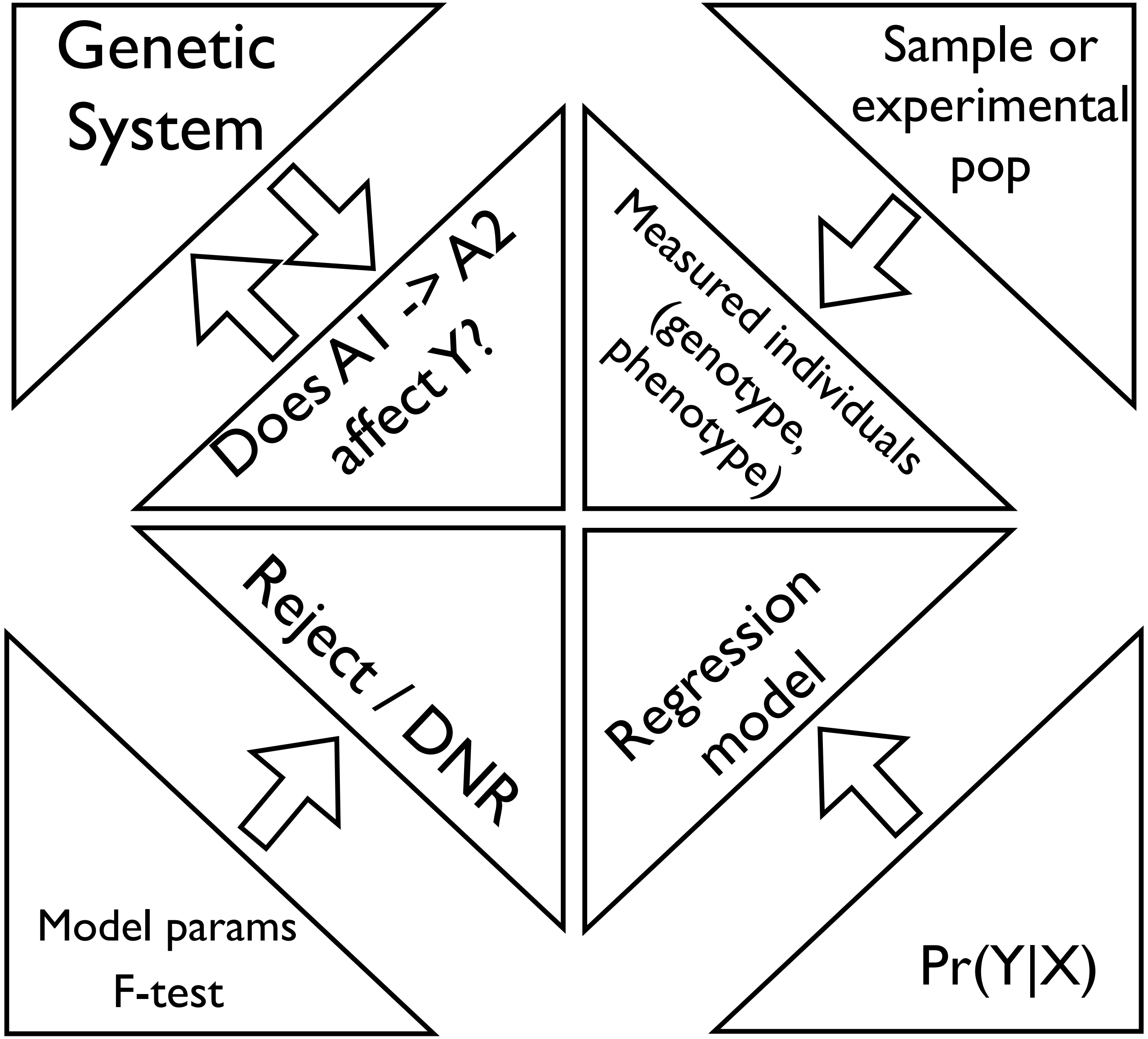
jgm45@cornell.edu

April 28, 2021 (W) 2:00-3:00PM

Summary of Optional Lecture 1

- Today we will discuss important concepts in Population Genetics helpful for understanding Linkage Disequilibrium
- And the related concept of Haplotype Testing in GWAS

Conceptual Overview



Genetic System

Sample or experimental pop

Does A1 -> A2 affect Y?

Measured individuals (genotype, phenotype)

Reject / DNR

Regression model

Model params

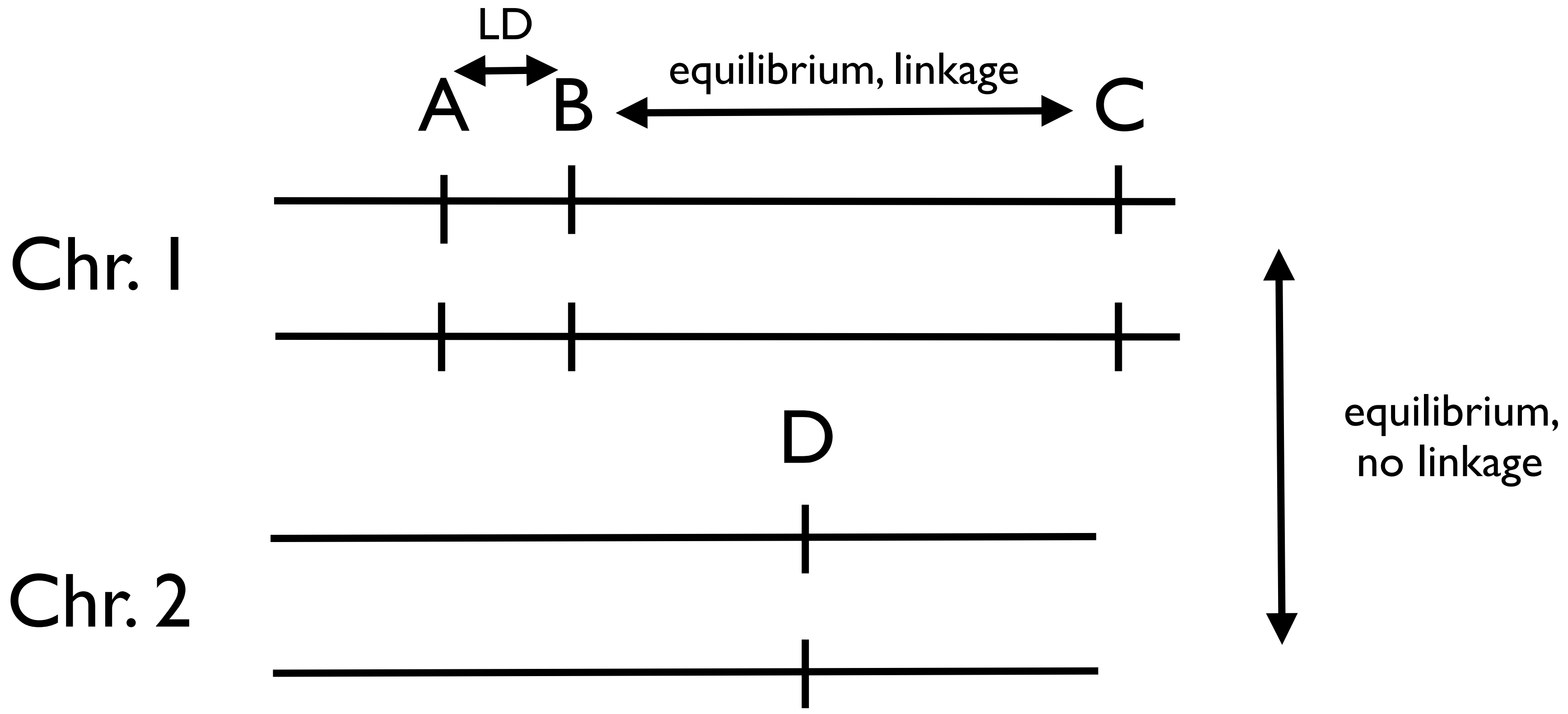
F-test

Pr(Y|X)

Linkage Disequilibrium (LD)

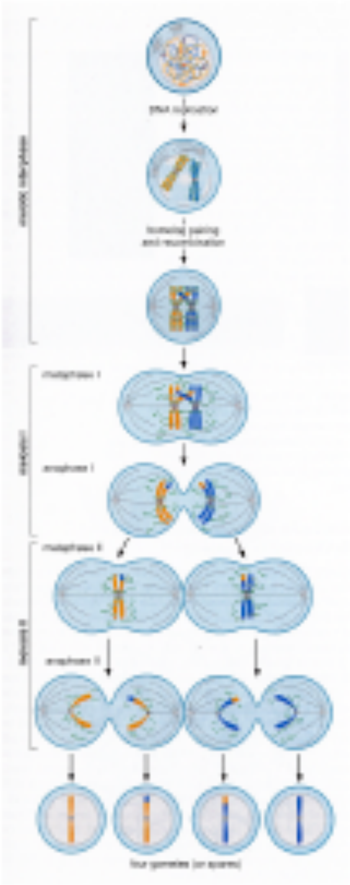
- Mapping the position of a causal polymorphism in a GWAS requires there to be LD for genotypes that are both physically linked and close to each other AND that markers that are either far apart or on different chromosomes to be in equilibrium

- Note that dis*equilibrium* includes both *linkage disequilibrium* AND other types of dis*equilibrium* (!!), e.g. gametic phase disequilibrium

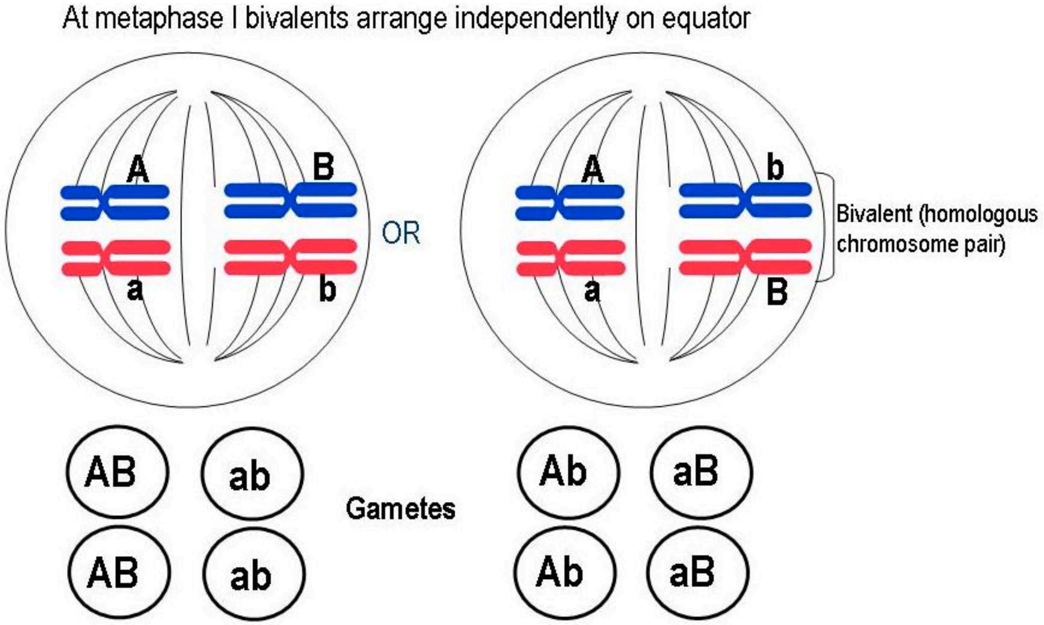


Different chromosomes I

- Polymorphisms on different chromosomes tend to be in equilibrium because of independent assortment and random mating, i.e. random matching of gametes to form zygotes



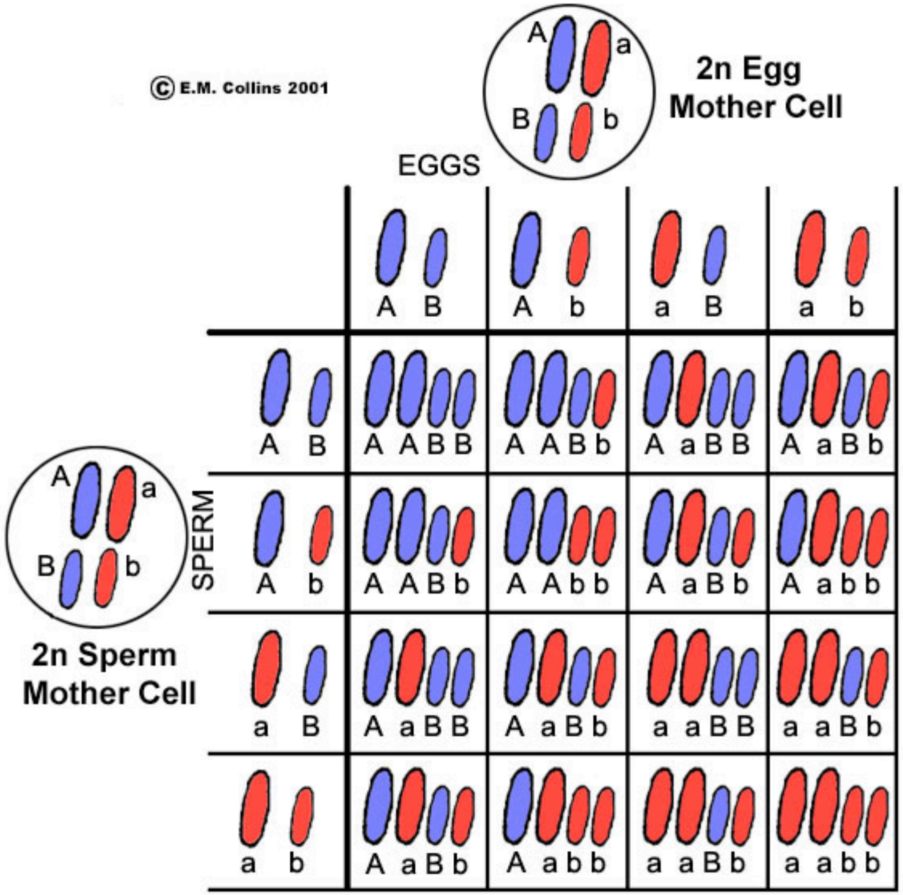
Credit: Watson et al., *Molecular Biology of the Gene*, CSHL Press, 2004


At metaphase I bivalents arrange independently on equator
A
B
a
b
OR
A
b
a
B
Bivalent (homologous chromosome pair)
AB
ab
AB
ab
Gametes
Ab
aB
Ab
aB

Copyright: http://geneticssuite.net/node/21

Different chromosomes II

- Polymorphisms on different chromosomes tend to be in equilibrium because of independent assortment and random mating, i.e. random matching of gametes to form zygotes


© E.M. Collins 2001

	A B	A b	a B	a b
A B	A A B B	A A B b	A a B B	A a B b
A b	A A B b	A A b b	A a B b	A a b b
a B	A a B B	A a B b	a a B B	a a B b
a b	A a B b	A a b b	a a B b	a a b b

Different chromosomes III

- More formally, we represent independent assortment as:

$$Pr(A_iB_k) = Pr(A_i)Pr(B_k)$$

- For random pairing of gametes to produce zygotes:

$$Pr(A_iB_k, A_jB_l) = Pr(A_iB_k)Pr(A_jB_l)$$

- Putting this together for random pairing of gametes to produce zygotes we get the conditions for equilibrium:

$$Pr(A_iB_k, A_jB_l) = Pr(A_iB_k)Pr(A_jB_l)$$

$$= Pr(A_i)Pr(A_j)Pr(B_k)Pr(B_l) = Pr(A_iA_j)Pr(B_kB_l)$$

$$\Rightarrow (Corr(X_{a,A}, X_{a,B}) = 0) \cap (Corr(X_{a,A}, X_{d,B}) = 0)$$

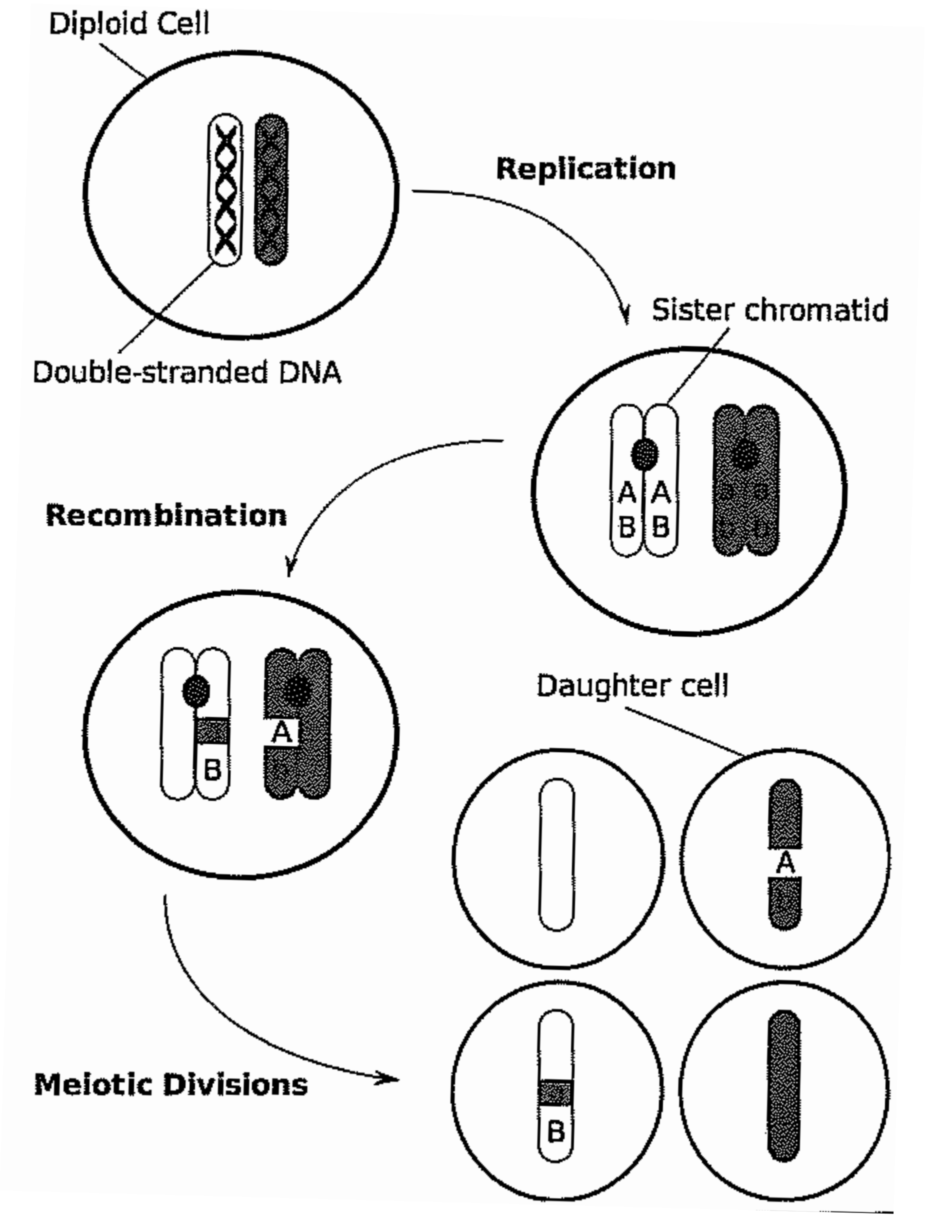
$$\cap (Corr(X_{d,A}, X_{a,B}) = 0) \cap (Corr(X_{d,A}, X_{d,B}) = 0)$$

Same chromosome I

- For polymorphisms on the same chromosome, they are linked so if they are in disequilibrium, they are in LD
- In general, polymorphisms that are closer together on a chromosome are in greater LD than polymorphisms that are further apart (exactly what we need for GWAS!)
- This is because of recombination, the biological process by which chromosomes exchange sections during meiosis
- Since recombination events occur at random throughout a chromosome (approximately!), the further apart two polymorphisms are, the greater the probability of a recombination event between them
- Since the more recombination events that occur between polymorphisms, the closer they get to equilibrium, this means markers closer together tend to be in greater LD

Same chromosome II

- In diploids, recombination occurs between pairs of chromosomes during meiosis (the formation of gametes)
- Note that this results in taking alleles that were physically linked on different chromosomes and physically linking them on the same chromosome



Same chromosome III

- To see how recombination events tend to increase equilibrium, consider an extreme example where alleles A1 and B1 always occur together on a chromosome and A2 and B2 always occur together on a chromosome:

$$Pr(A_1B_2) = 0, \; Pr(A_2B_1) = 0$$

$$Corr(X_{a,A}, X_{a,B}) = 1 \text{ AND } Corr(X_{d,A}, X_{d,B}) = 1$$

- If there is a recombination event, most chromosomes are A1-B1 and A2-B2 but now there is an A1-B2 and A2-B1 chromosome such that:

$$Pr(A_1B_2) \neq 0, \; Pr(A_2B_1) \neq 0$$

$$Corr(X_{a,A}, X_{a,B}) \neq 1 \text{ AND } Corr(X_{d,A}, X_{d,B}) \neq 1$$

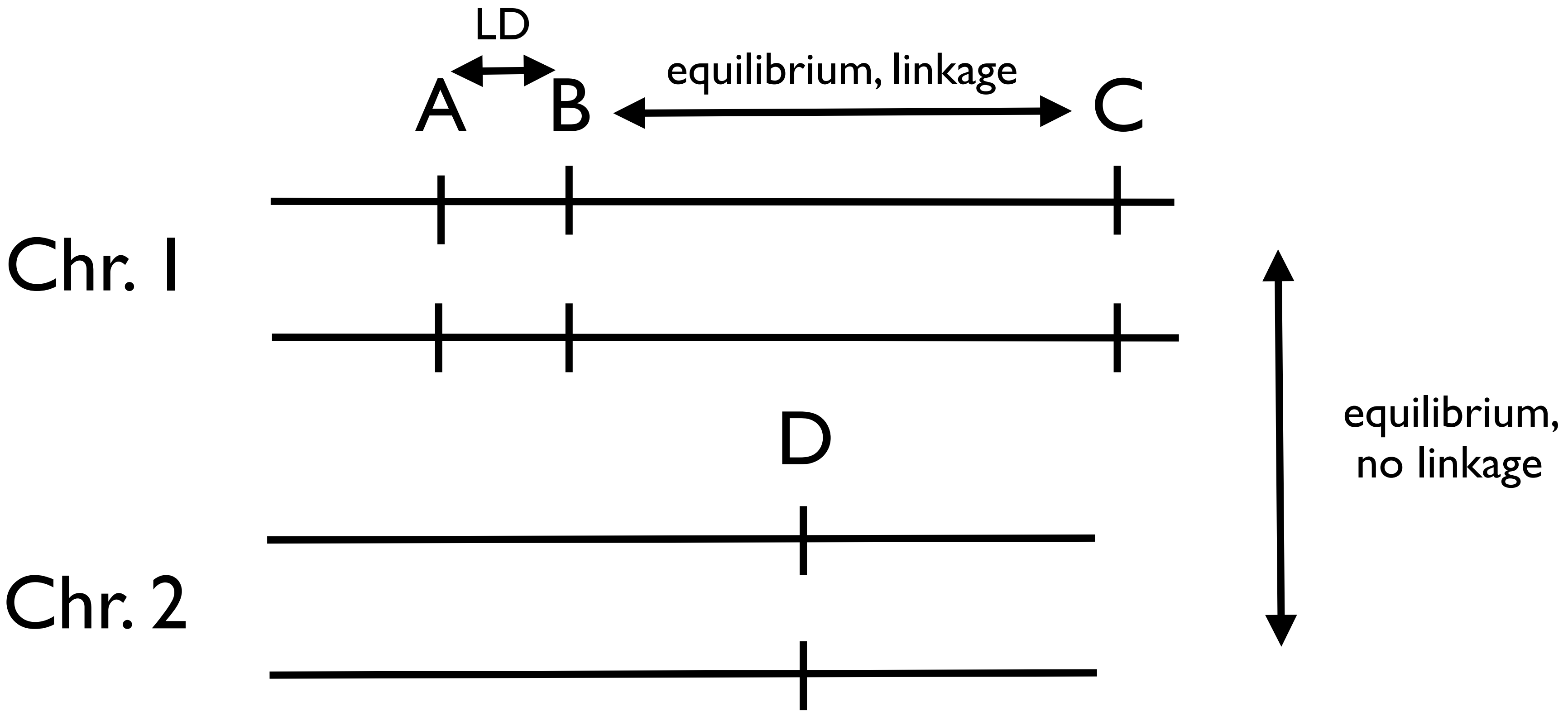
- Note recombination events disproportionally lower the probabilities of the more frequent pairs!

- This means over time, the polymorphisms will tend to increase equilibrium (decrease LD)

- Since the more recombination events, the greater the equilibrium, polymorphisms that are further apart will tend to be in greater equilibrium, those closer together in greater LD

Linkage Disequilibrium (LD)

- Mapping the position of a causal polymorphism in a GWAS requires there to be LD for genotypes that are both physically linked and close to each other AND that markers that are either far apart or on different chromosomes to be in equilibrium
- Note that dis*equilibrium* includes both *linkage disequilibrium* AND other types of dis*equilibrium* (!!), e.g. gametic phase disequilibrium



Side topic: connection coin flip models to allele / genotypes

- Recall we the one coin flip example (how does the parameter of Bernoulli relate to MAF?):

$$\Omega = \{H, T\} \qquad X(H) = 0, X(T) = 1$$

$$Pr(X = x|p) = P_X(x|p) = p^x(1-p)^{1-x}$$

- The following model for two coin flips maps perfectly on to the model of genotypes (e.g., represented as number of A1 alleles) under Hardy-Weinberg equilibrium (e.g., for MAF = 0.5):

$$X(HH) = 0, X(HT) = 1, X(TH) = 1, X(TT) = 2$$

$$Pr(HH) = Pr(HT) = Pr(TH) = Pr(TT) = 0.25$$

$$P_X(x) = Pr(X = x) = \begin{cases} Pr(X = 0) = 0.25 \\ Pr(X = 1) = 0.5 \\ Pr(X = 2) = 0.25 \end{cases} \quad Pr(X = x|n,p) = P_X(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x}$$

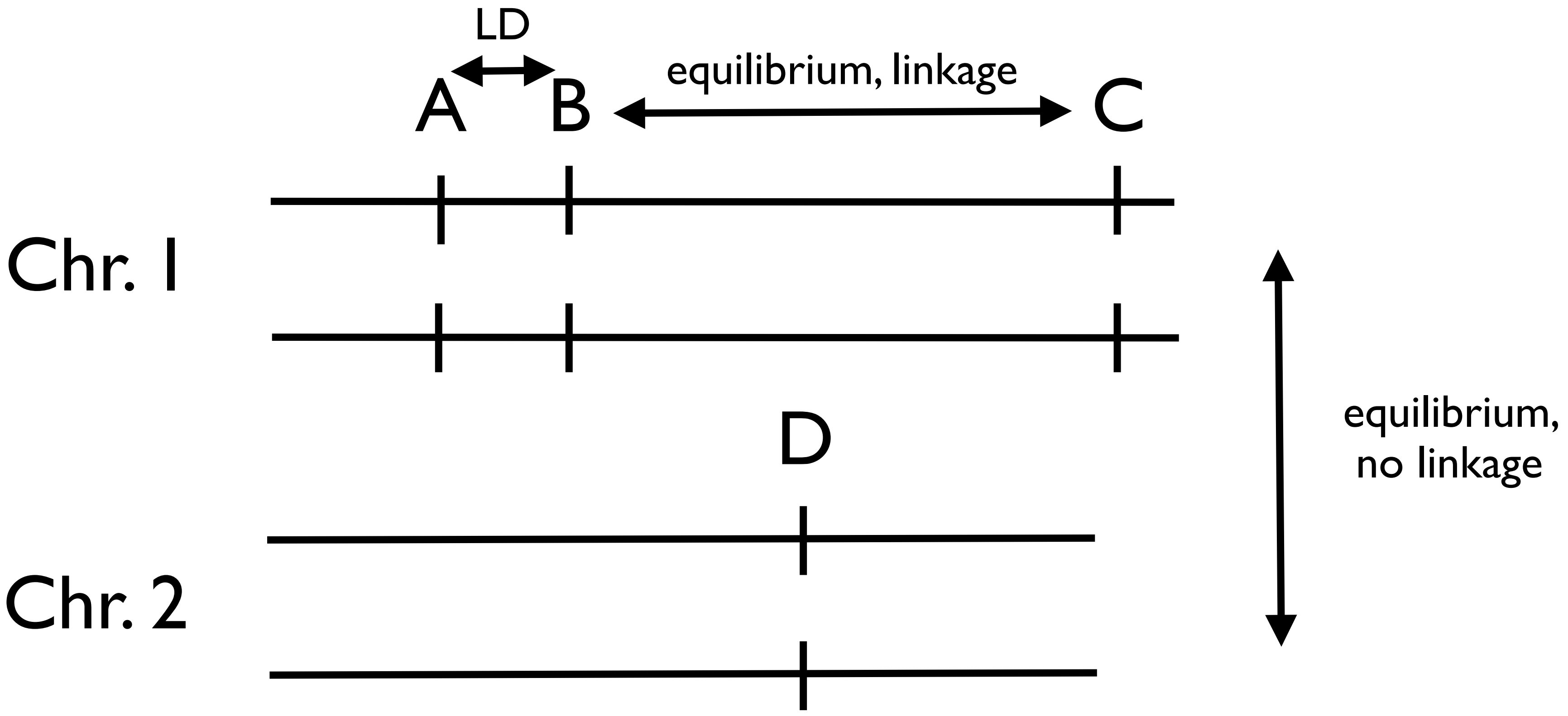
- Note that the model need not conform to H-W since consider the following model (we could use a multinomial probability distribution):

$$Pr(X_1 = 0, X_2 = 0) = 0.0, Pr(X_1 = 0, X_2 = 1) = 0.25$$
$$Pr(X_1 = 1, X_2 = 0) = 0.25, Pr(X_1 = 1, X_2 = 1) = 0.25$$
$$Pr(X_1 = 2, X_2 = 0) = 0.25, Pr(X_1 = 2, X_2 = 1) = 0.0$$

Review: Linkage Disequilibrium (LD)

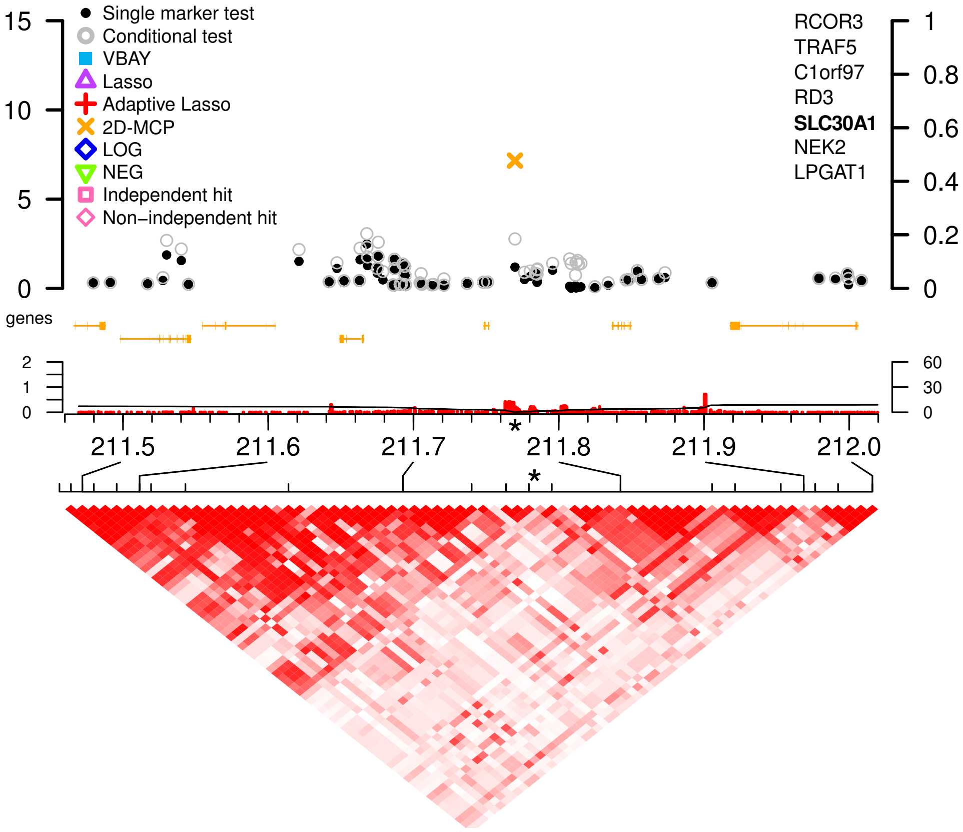
- Mapping the position of a causal polymorphism in a GWAS requires there to be LD for genotypes that are both physically linked and close to each other AND that markers that are either far apart or on different chromosomes to be in equilibrium

- Note that dis*equilibrium* includes both *linkage disequilibrium* AND other types of dis*equilibrium* (!!), e.g. gametic phase disequilibrium



Patterns and representing LD

- We often see LD among a set of contiguous markers, using either r-squared or D', with the "triangle, half-correlation matrices" where darker squares indicating higher LD (values of these statistics, e.g. LD in a "zoom-in" plot:



Measuring LD I

- There are *many* statistics used to represent LD but we will present the two most common
- For the first, define the correlation:

$$r = \frac{Pr(A_i, B_k) - Pr(A_i)Pr(B_k)}{\sqrt{Pr(A_i)(1 - Pr(A_i)}\sqrt{Pr(B_k)(1 - Pr(B_k)}}$$

- As a measure of LD, we will consider this squared:

$$r^2 = \frac{(Pr(A_i, B_k) - Pr(A_i)Pr(B_k))^2}{(Pr(A_i)(1 - Pr(A_i))(Pr(B_k)(1 - Pr(B_k))}$$

- Note that this is always between one and zero!

Phasing

- To get a sense of the phasing problem, consider a case where we have two markers that are right next to each other on a chromosome and we know we want to put them together in a haplotype block

- Say one marker is (A,T) and the other marker is (G,C) and we are considering a diploid individual who is a heterozygote for both of these markers, which of the marker alleles are physically linked in this individual?

- Figuring this out for individuals in a sample is the phasing problem and there are many algorithms for accomplishing this goal (note that in the future, technology may make this a non-issue...)

Measuring LD II

- A “problem” with r-squared is that when the MAF of *A* or *B* is small, this statistic is small
- For the second measure of LD, we will define a measure D’ that is not as dependent on MAF:

$$D = Pr(A_i, B_k) - Pr(A_i)Pr(B_k)$$

$$D' = \frac{D}{min(Pr(A_1 B_2), Pr(A_2, B_1))} \text{if} D > 0$$

$$D' = \frac{D}{min(Pr(A_1 B_1), Pr(A_2, B_2))} \text{if} D < 0$$

- Note that this is always between -1 and 1 (!!)

Haplotype testing I

- We have just extended our GWAS framework to make use of LD in a different manner than we have with our basic GWAS testing approach

- In this case, let’s consider using *haplotype* alleles in our testing framework

- Note that a haplotype collapses genetic marker information but in some cases, testing using haplotypes is more effective than testing one genetic marker at a time

Haplotype testing II

- **Haplotype -** a series of ordered, linked alleles that are inherited together
- For the moment, let's consider a haplotype to define a "function" that takes a set of alleles at several loci A, B, C, D, etc. and outputs a haplotype allele:

$$h = f(A_i, B_j, ...)$$

- For example, if these loci are each a SNP with the following alleles (A,G), (A,T),(G,C),(G,C) we could define the following haplotype alleles:

$$h_1 = (A, A, C, C) \qquad h_2 = (G, T, G, G)$$

Haplotype testing III

- Note that how we define haplotype alleles is somewhat arbitrary but in general, we define a haplotype for a set of genetic markers (loci) that are physically linked that are frequently occur in a population

- How many markers is somewhat arbitrary, e.g. we often define sets that match observed patterns of LD

- How many haplotype alleles we define is also somewhat arbitrary, where we define haplotype alleles that have appreciable frequenecy in the population

  - For example, four the four loci with alleles (A,G), (A,T),(G,C),(G,C) how many haplotype alleles could we define?

  - However, it could be that only the following two combinations have relatively "high" allele frequencies (say >0.05 = arbitrary!)

$$h_1 = (A, A, C, C) \qquad h_2 = (G, T, G, G)$$

  - In such a case, we can collapse the many alleles into just a few!

Haplotype testing IV

- As an example of haplotype allele collapsing, say for our case of four loci (A,G), (A,T),(G,C),(G,C), we have lots of LD (!!) such that there are only 4 alleles in the population (i.e. all other combinations have frequency of zero!):

$$h_1^* = (A, A, C, C), h_2^* = (G, T, G, G), h_3^* = (A, A, G, C), h_4^* = (G, T, C, G)$$

- Let's also say that the frequencies of the third and fourth of these in the population are < 0.01

- In this case, we can define just two haplotype alleles that collapse the other alleles as follows (where * means "any" genetic marker allele):

$$h_1 = (A, A, *, C) \qquad h_2 = (G, T, *, G)$$

$$h_1 = h_1^* \cup h_3^* \qquad h_2 = h_2^* \cup h_4^*$$

- NOTE: we are therefore loosing information using this approach!!

GWAS with haplotypes I

- Once we have defined haplotype alleles, we can proceed with a GWAS using our framework (just substitute haplotype alleles and genotypes for genetic marker alleles and genotypes!)
- For example, in a case where we only have two haplotype alleles, we can code our independent variables for our regression model as follows:

$$X_a(h_1h_1) = -1, X_a(h_1h_2) = 0, X_a(h_2h_2) = 1$$

$$X_d(h_1h_1) = -1, X_d(h_1h_2) = 1, X_d(h_2h_2) = -1$$

- All other aspects remain the same (although what is the effect on our interpretation of where the causal polymorphism is located?)

GWAS with haplotypes II

- Given that we are losing information by using a haplotype testing approach in a GWAS, why might we want to use this approach?

- As one example consider the following case of haplotypes in a population:

$$\begin{array}{ccccc}
A_1 & B_1 & (C_1)* & D_2 & E_1 \\
A_1 & B_2 & (C_1)* & D_1 & E_1 \\
A_2 & B_1 & (C_1)* & D_1 & E_1 \\
A_1 & B_1 & (C_1)* & D_1 & E_2 \\
A_2 & B_2 & (C_2)* & D_1 & E_2 \\
A_2 & B_1 & (C_2)* & D_2 & E_2 \\
A_1 & B_2 & (C_2)* & D_2 & E_2 \\
A_2 & B_2 & (C_2)* & D_2 & E_1
\end{array}$$

Advantages of haplotype testing

- In some cases (system and sample dependent!), the haplotype is a better “tag” of the causal polymorphism than any of the surrounding markers
- In such a case, the $Corr(X_h, X) > Corr(X', X)$ and therefore has a higher probability of correctly rejecting the null hypothesis
- Another “advantage” is by putting together markers, we are performing less total tests in our GWAS (in what sense is this an advantage!?)

Disadvantages of haplotype testing

- Collapsing to haplotypes may produce a better tag but it also may not (!!), i.e. sometimes (in fact often!) individual genetic markers are better tags of the causal polymorphism

- Another disadvantage is resolution, since we absolutely cannot resolve the position of the causal polymorphism to a position smaller than the range of the haplotype alleles, i.e. large haplotypes can have smaller resolution

- If we had measured the causal polymorphism in our data, should we use haplotype testing (i.e. in the future, the importance of haplotype testing may decrease)

Should I apply haplotype testing in my GWAS?

- Yes! but apply both an individual marker testing approach (always!) as well as a haplotype test (optional)
- The reason is that we never know the true answer in our GWAS (as with any statistical analysis!) so it doesn't hurt us to explore our dataset with as many techniques as we want to apply
- In fact, this will be a continuing theme of the class, i.e. keep analyzing GWAS with as many methods as you find useful
- However, since we never know the right answer for certain, if we get conflicting results, which one do we interpret as "correct"!?

Where do haplotypes come from?

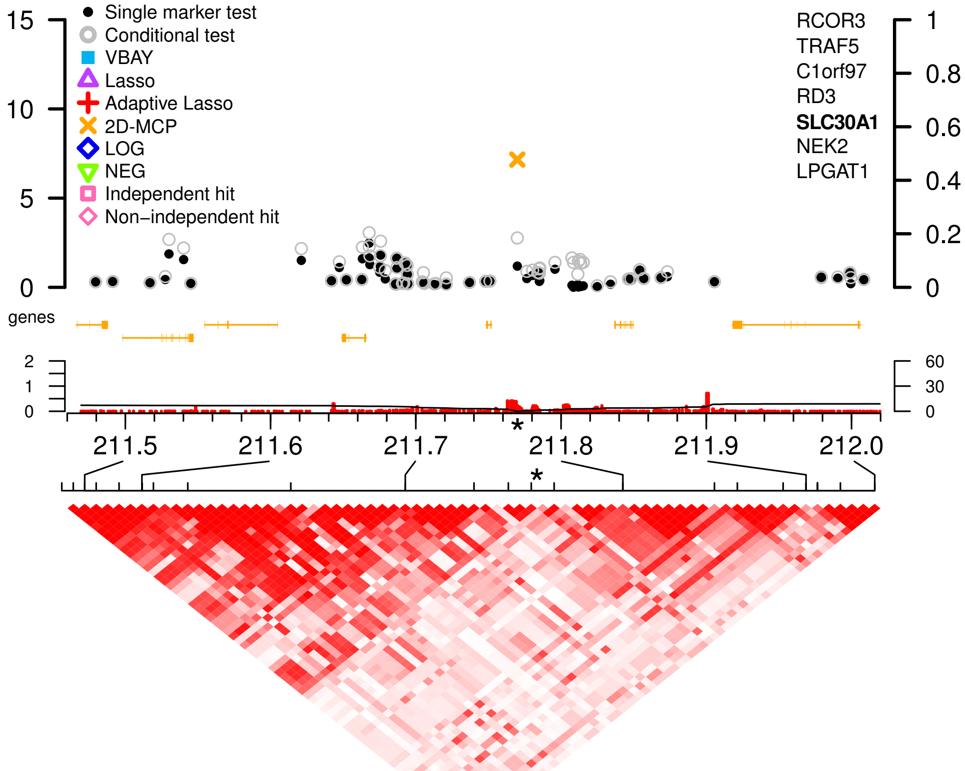
- A deep discussion of the origin of haplotypes (remember: a fuzzy definition!) is another subject that is in the realm of population genetics and therefore we cannot discuss this in detail in this class (again: I encourage you to take a class on population genetics!)
- However, we can get an intuition about where haplotypes come from by remembering that the origin of new haplotype alleles are *mutations* and that new haplotype alleles can be produced by *recombination*
- In fact, these two processes also underlie the amount of LD in the population and therefore what blocks of alleles are inherited as a haplotype (and we therefore use them to define haplotypes using system specific criteria)

Defining haplotypes

- We could spend multiple lectures on how people define haplotypes for given systems and the algorithms used for this purpose (so we will just briefly mention the main concepts here)
- To define haplotypes, we need to "phase" measured genotype markers, decide on the number of genotype markers to put together into a haplotype block, and decide how many haplotype alleles to consider
- Remember: there are no universal rules for doing this (system dependent!)

Deciding on how many genotypes to include in a haplotype block

- Again, while there is no set rule, how we decide on genotypes to include in a haplotype block depends on LD
- The general rule: if we have a set of markers in high LD with each other but low LD with other markers, we use this as a guide for defining the haplotype block



Deciding on how many haplotype alleles to consider

- Again, there are no set rules for how many haplotype alleles to define, but in general, we define a set where the frequency in a population is above some MAF threshold (which depends on the system)
- With a MAF cutoff of say 0.05, this generally limits us to 2-5 haplotype alleles (e.g. in humans!)
- There are however cases where we might want to consider rarer haplotypes (what are some of these?)

Haplotype GWAS wrap-up

- Haplotypes are a physical and sampling consequence of how genetic systems work (just like LD!)
- Definitions of haplotype blocks and haplotype alleles depend on the system and context (fuzzy definition)
- Regardless of how we define them, once we have haplotype alleles, we can use them as we would genetic markers in our GWAS analysis framework
- While optional, it is never a bad idea to perform a haplotype analysis of your GWAS in addition to your single marker analysis (ALWAYS do a single marker analysis)

That’s it for today

- See you next time!