Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

Lecture 16: Introduction to GWAS

Jason Mezey
March 21, 2023 (T) 8:05-9:20

Announcements

- Homework #4
  - Due 11:59pm March 27
  - We will cover the last topics you need for QUESTION #2 during this lecture and next (!!)
- Next week you will have you Midterm Exam (!!)
  - Available Weds (March 29)
  - Due 11:59pm Fri (March 31)
  - If you prepare ahead of time (!!) it should only take you a few hours to complete!

Quantitative Genomics and Genetics - Spring 2023
BTRY 4830/6830; PBSB 5201.01

Midterm Exam

Available on CMS by 11AM (ET), Weds., March 29
Due 11:59PM (ET) Fri., March 31

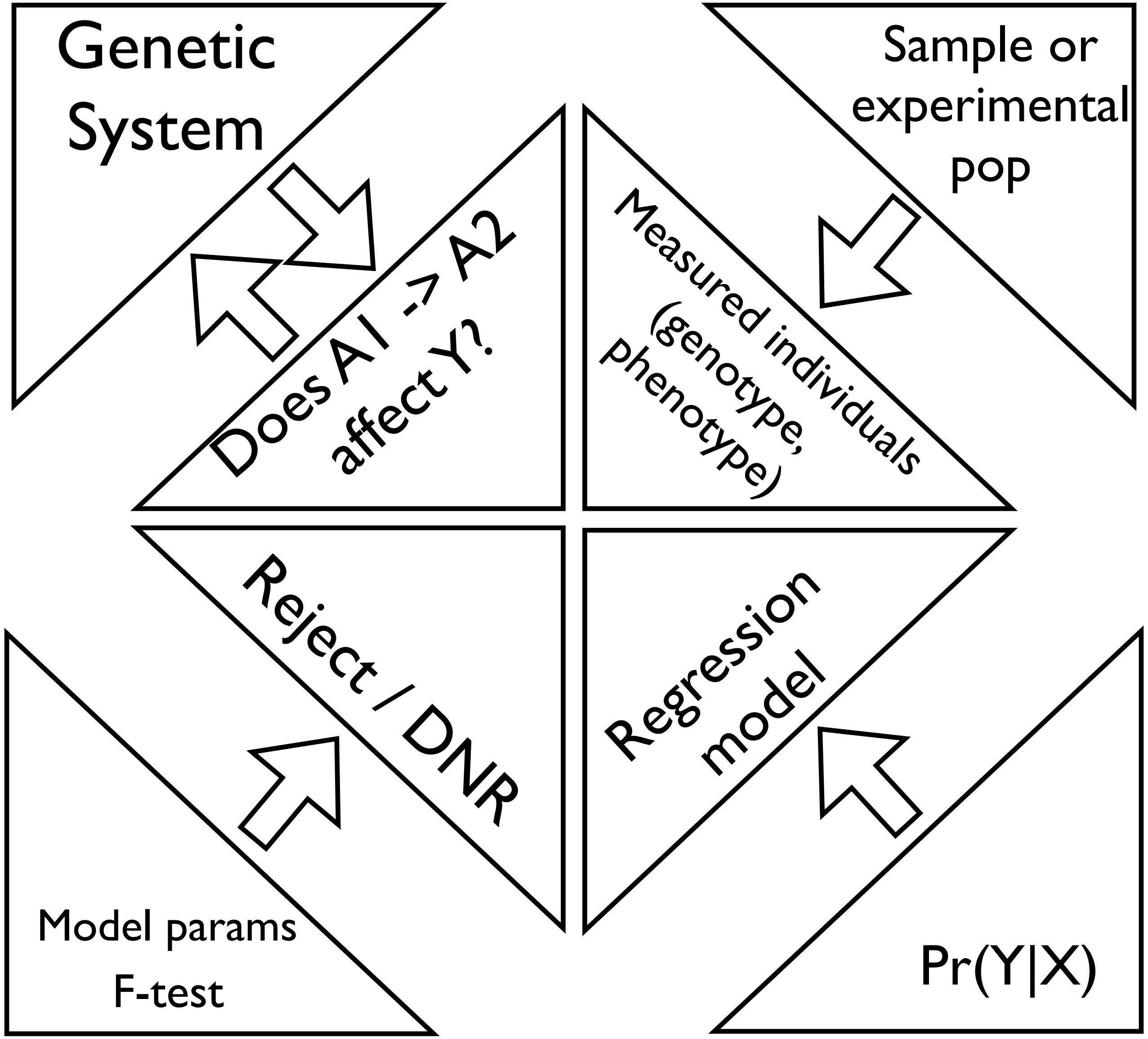
PLEASE NOTE THE FOLLOWING INSTRUCTIONS:

1. You are to complete this exam alone. The exam is open book, so you are allowed to use any books or information available online, your own notes and your previously constructed code, etc. **HOWEVER YOU ARE NOT ALLOWED TO COMMUNICATE OR IN ANY WAY ASK ANYONE FOR ASSISTANCE WITH THIS EXAM IN ANY FORM e.g., DO NOT POST PUBLIC MESSAGES ON PIAZZA!** (the only exceptions are Mitch, Sam, and Dr. Mezey, e.g., you MAY send us a private message on PIAZZA). As a non-exhaustive list this includes asking classmates or ANYONE else for advice or where to look for answers concerning problems, you are not allowed to ask anyone for access to their notes or to even look at their code whether constructed before the exam or not, etc. You are therefore only allowed to look at your own materials and materials you can access on your own. In short, work on your own! Please note that you will be violating Cornell's honor code if you act otherwise.

Summary of lecture 16: Introduction to GWAS

- Last lecture, we completed our discuss of genetic inference for a single position in the genome (e.g., SNP)
- Today, we will begin discuss GWAS!

Conceptual Overview



Review: Genetic system

- **causal mutation -** a position in the genome where an experimental manipulation of the DNA would produce an effect on the phenotype under specifiable conditions

- Formally, we may represent this as follows:

$$A_1 \rightarrow A_2 \Rightarrow \Delta Y | Z$$

- Note: that this definition considers "under specifiable" conditions" so the change in genome need not cause a difference under every manipulation (just under broadly specifiable conditions)

- Also note the symmetry of the relationship

- Identifying these is the core of quantitative genetics/genomics (why do we want to do this!?)

- What is the perfect experiment?

- Our experiment will be a statistical experiment (sample and inference!)

Review: Genetic estimation

- Let's look at the structure of this estimator:

$$\mathbf{y} = \mathbf{x}\beta + \epsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,a} & x_{1,d} \\ 1 & x_{2,a} & x_{2,d} \\ \vdots & \vdots & \ddots \\ 1 & x_{n,a} & x_{n,d} \end{bmatrix} \begin{bmatrix} \beta_\mu \\ \beta_a \\ \beta_d \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$MLE(\hat{\beta}) = (\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathrm{T}}\mathbf{y}$$

$$MLE(\hat{\beta}) = \begin{bmatrix} \hat{\beta}_\mu \\ \hat{\beta}_a \\ \hat{\beta}_d \end{bmatrix}$$

Review Genetic hypothesis testing

- We are going to test the following hypothesis:

$$H_0 : \beta_a = 0 \cap \beta_d = 0$$

$$H_A : \beta_a \neq 0 \cup \beta_d \neq 0$$

- To do this, we need to construct the following test statistic (for which we know the distribution!):

$$T(\mathbf{y}, \mathbf{x}_a, \mathbf{x}_d | H_0 : \beta_a = 0 \cap \beta_d = 0)$$

- Specifically, we are going to construct a likelihood ratio test (LRT)
- This is calculated using the same structure that we have discussed (i.e. ratio of likelihoods that take values of parameters maximized under the null and alternative hypothesis)
- In the case of a regression (not all cases!) we can write the form of the LRT for our null in an alternative (but equivalent!) form
- In addition, our LRT has an exact distribution for all sample sizes *n* (!!)

Review: Genetic hypothesis testing

- To construct our LRT for our null, we will need several components, first the predicted value of the phenotype for each individual:

$$\hat{y}_i = \hat{\beta}_\mu + x_{i,a}\hat{\beta}_a + x_{i,d}\hat{\beta}_d$$

- Second, we need the "Sum of Squares of the Model" (SSM) and the "Sum of Squares of the Error" (SSE):

$$SSM = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \qquad SSE = \sum_{n=1}^{n}(y_i - \hat{y}_i)^2$$

- Third, we need the "Mean Squared Model" (MSM) and the "Mean Square Error" (MSE) with degrees of freedom (df) $df(M) = 3 - 1 = 2$ and : $df(E) = n - 3$

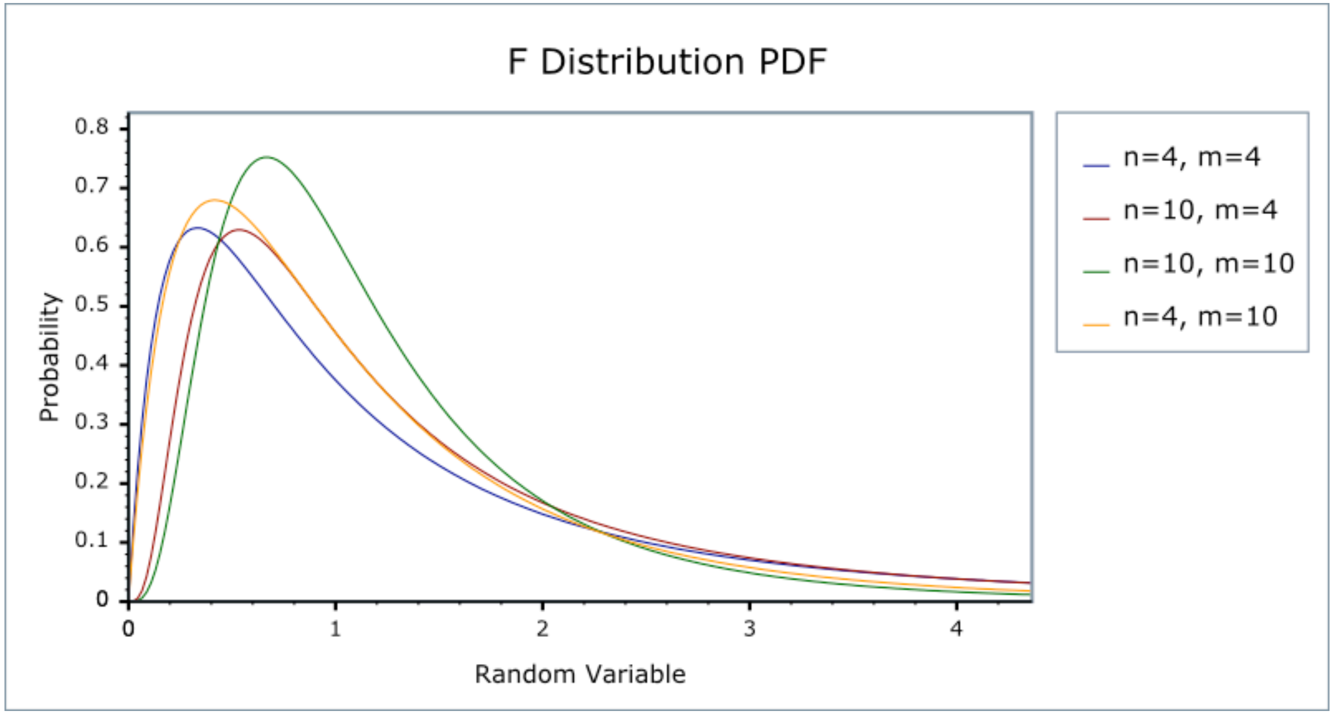
$$MSM = \frac{SSM}{df(M)} = \frac{SSM}{2} \qquad MSE = \frac{SSE}{df(E)} = \frac{SSE}{n-3}$$

- Finally, we calculate our (LRT!) statistic, the F-statistic with degrees of freedom [2, n-3]:

$$F_{[2,n-3]} = \frac{MSM}{MSE}$$

Review: Genetic hypothesis testing

- In general, the F-distribution (continuous random variable!) under the H0 has variable forms that depend on d.f.:



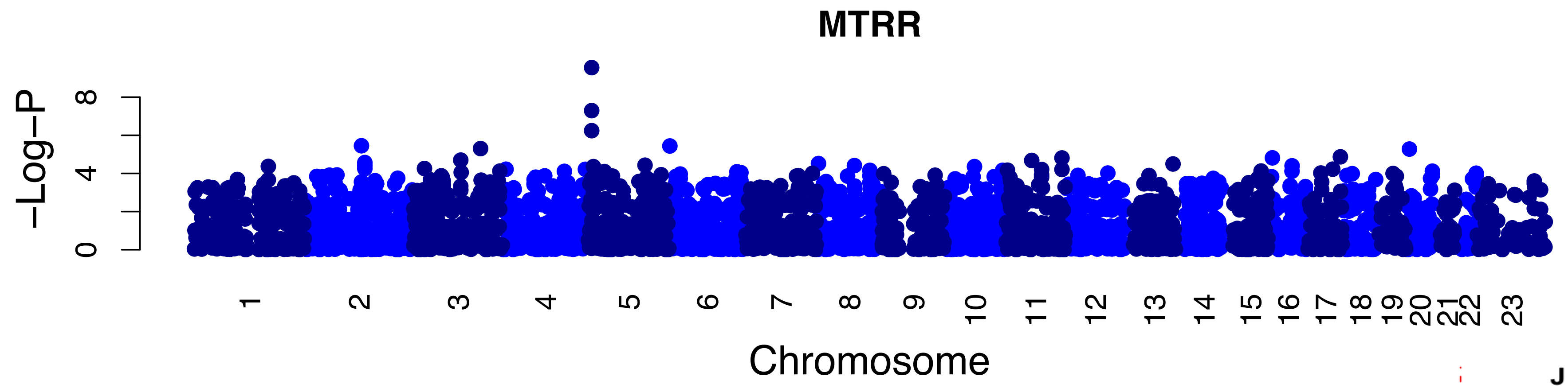
- Note when calculating a p-value for the genetic model, we consider the value of the F-statistic we observe or more extreme towards positive infinite (!!) using the F-distribution with [2,n=3] d.f.

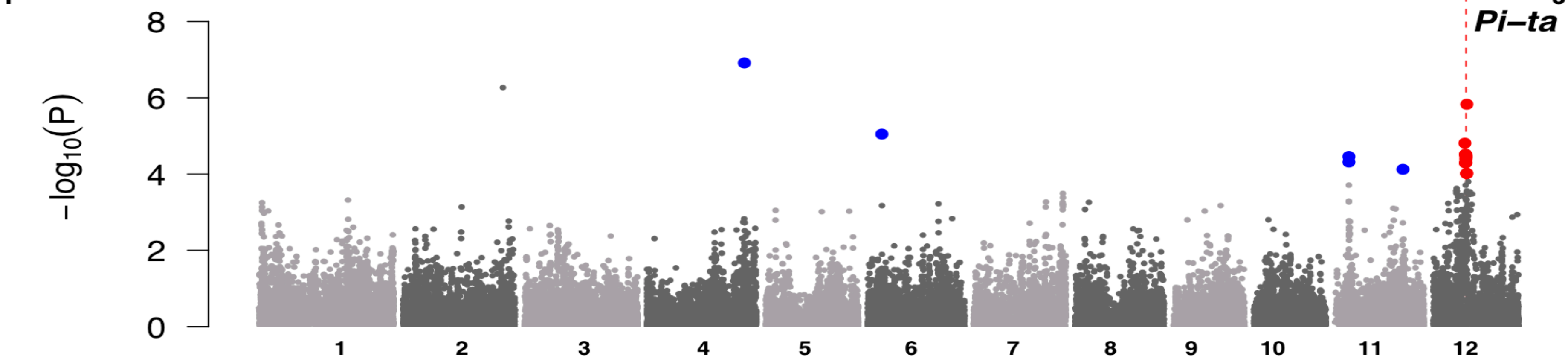
- However, also this is actually a two-tailed test (what is going on here (!?)

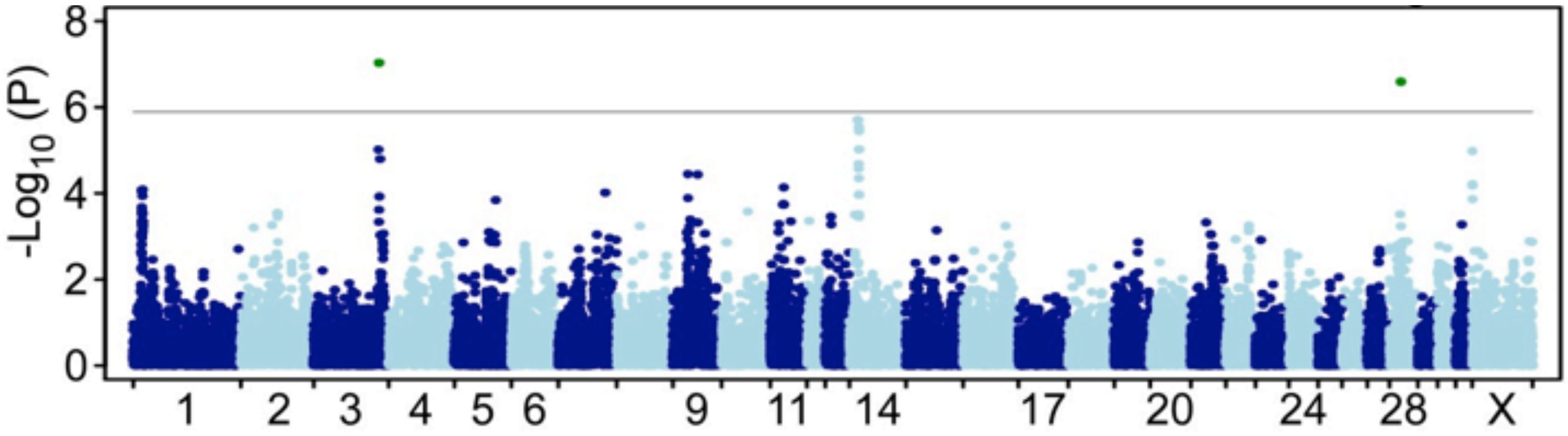
Review: Quantitative genomic analysis I

- We now know how to assess the null hypothesis as to whether a polymorphism has a causal effect on our phenotype
- Occasionally we will assess this hypothesis for a single genotype
- In quantitative genomics, we generally do not know the location of causal polymorphisms in the genome
- We therefore perform a hypothesis test of *many genotypes throughout the genome*
- This is a genome-wide association study (GWAS)

The Manhattan plot: examples





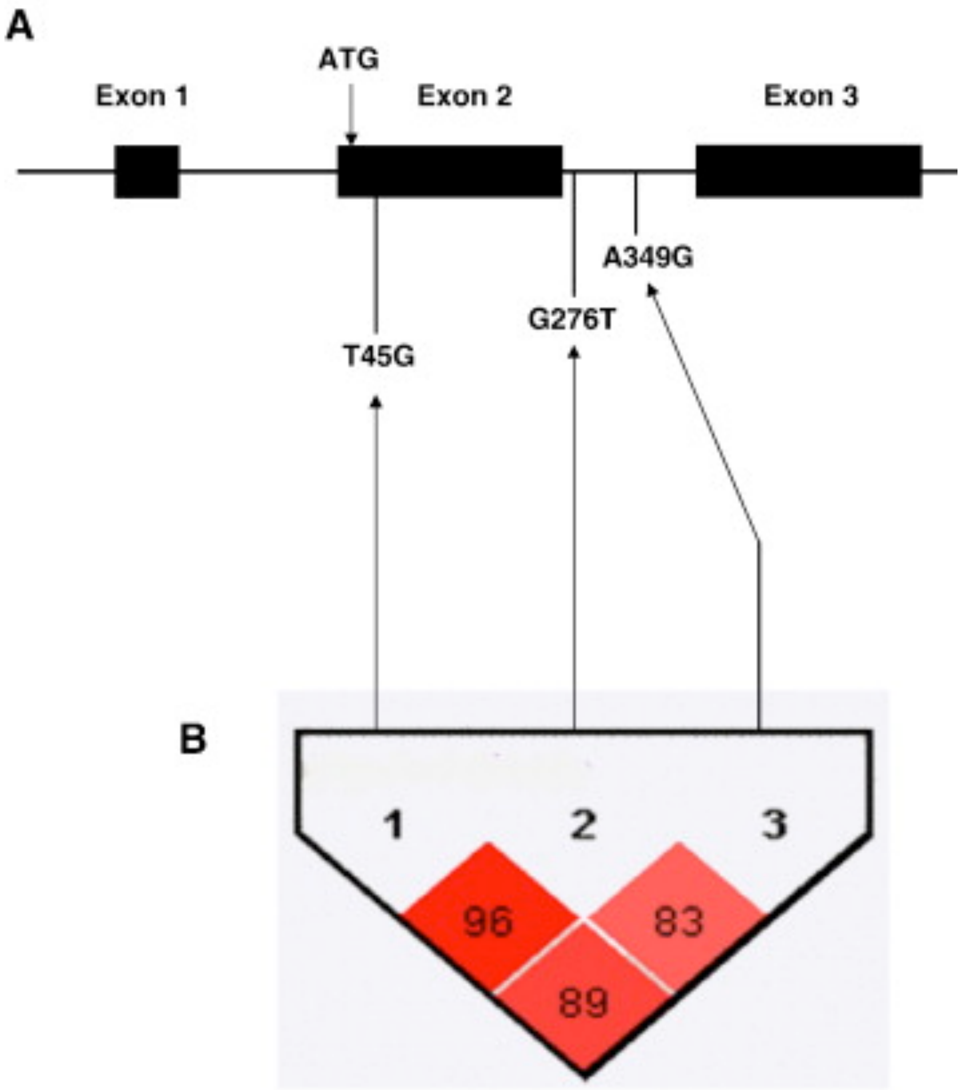


Review: Quantitative genomic analysis II

- Analysis in a GWAS raises (at least) two issues we have not yet encountered:
  - An analysis will consist of many hypothesis tests (not just one)
  - We often do not test the causal polymorphism (usually)
- Note that this latter issue is a bit strange (!?) - how do we assess causal polymorphisms if we have not measured the causal polymorphism?
- Also note that causal genotypes will begin to be measured in our GWAS with next-generation sequencing data (but the issue will still be present!)

Review: Correlation among genotypes

- If we test a (non-causal) genotype that is correlated with the causal genotype AND if correlated genotypes are in the same position in the genome THEN we can identify the genomic position of the casual genotype (!!)
- This is the case in genetic systems (why!?)
- Do we know which genotype is causal in this scenario?

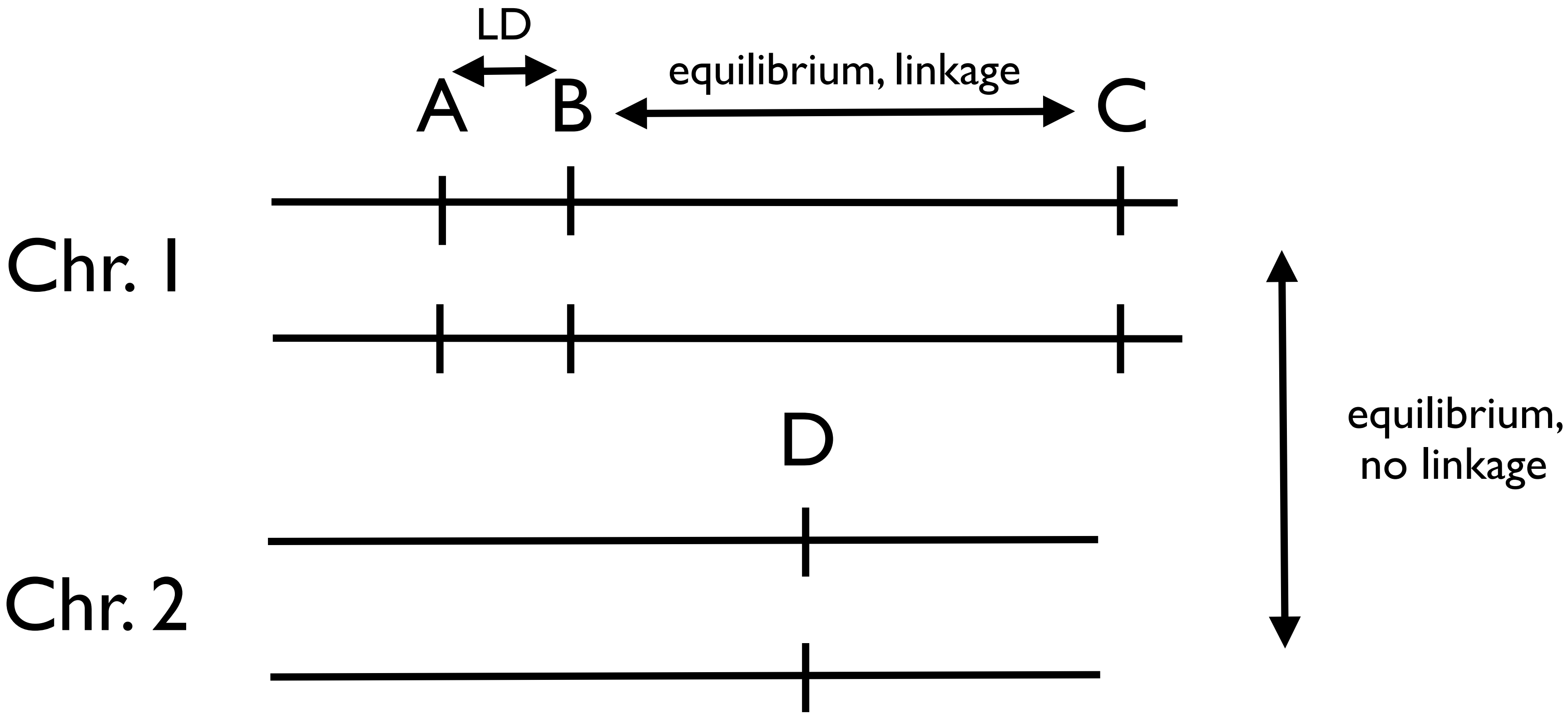


Copyright: Journal of Diabetes and its Complications; Science Direct; Vendramini et al

Linkage Disequilibrium

- 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

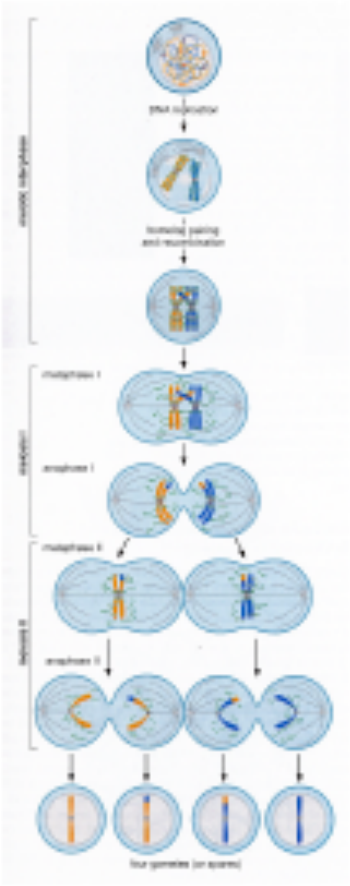


Population Genetic Causes of LD (in human populations)

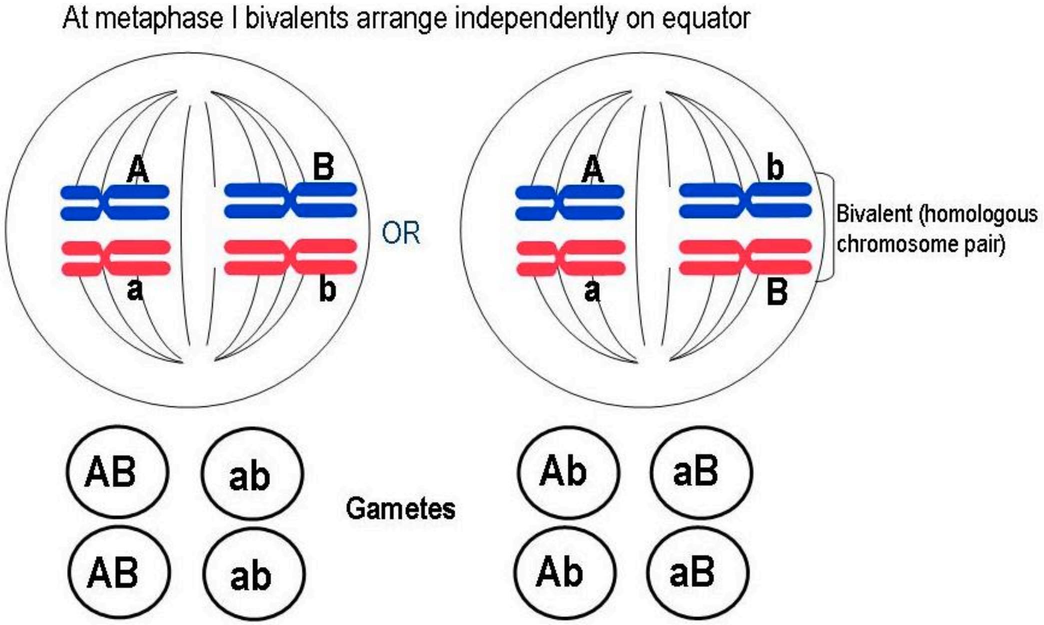
- The major factors responsible for patterns of LD in human populations are:
  - (1) Independent assortment of chromosomes
  - (2) “Random” mating
  - (3) Recombination
- Note (!!): this is the answer considering EXISTING variation in a population and therefore no MUTATION or MIGRATION
- Note that these factors explain LD in many other populations as well but there can be differences that lead to different patterns of LD (e.g., in natural populations, in breeding populations etc.)

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



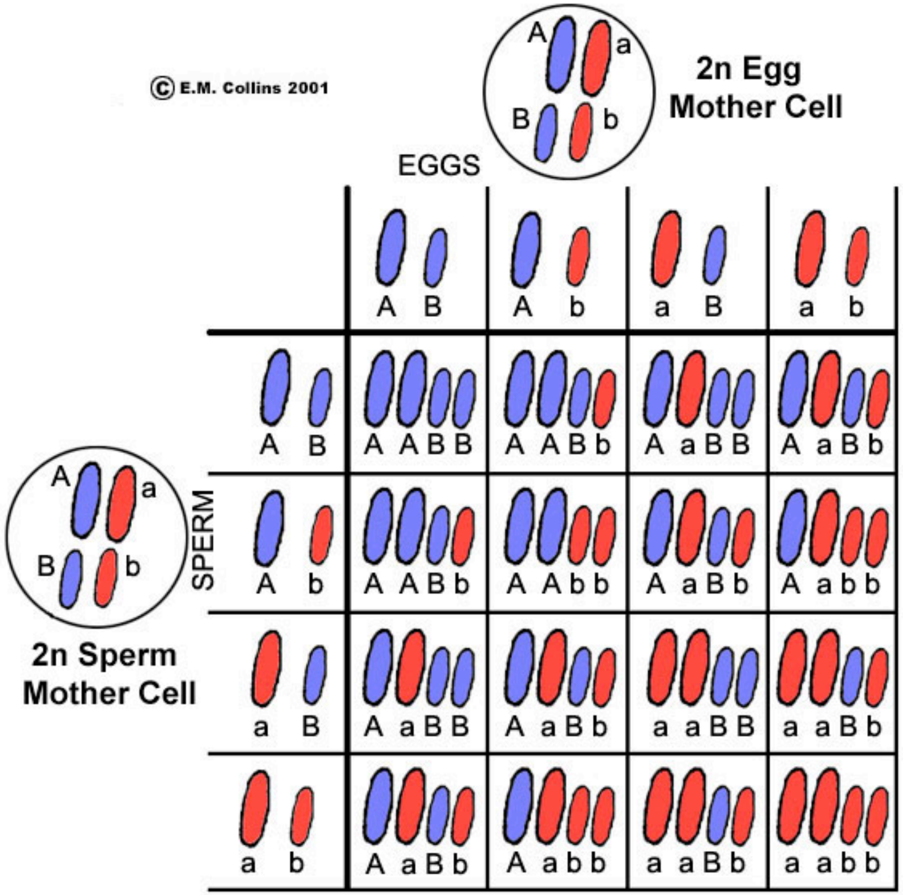
Credit: Watson et al., *Molecular Biology of the Gene*, CSHL Press, 2004


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



	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:

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

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

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

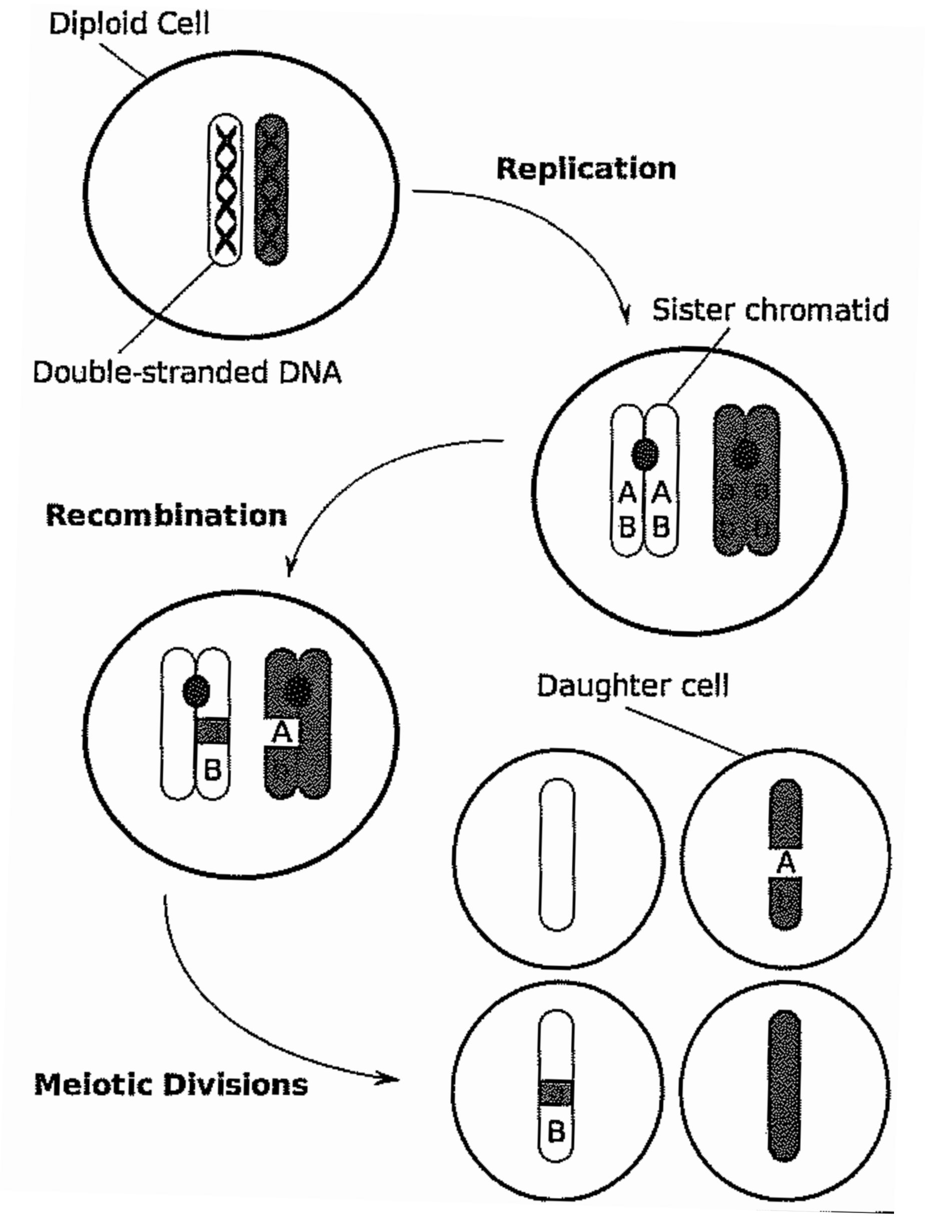
$$\cap(Corr(X_{d,A}, X_{a,B}) = 0) \cap (Corr(X_{d,A}, X_{d,B}) = 0)$$

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!

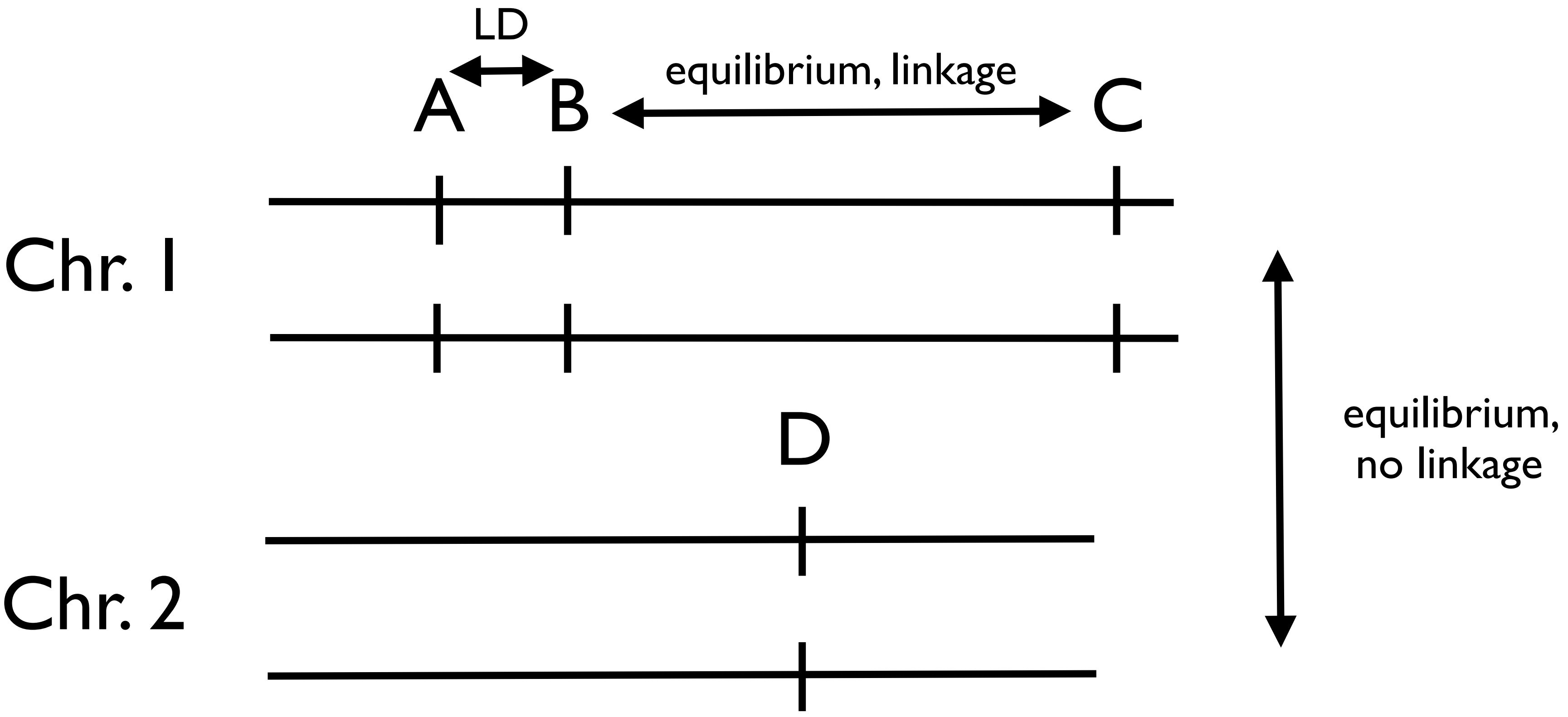
- This means over time, the polymorphisms will tend to increase equilibrium (decrease LD)

- 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):

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

Genome-Wide Association Study (GWAS)

- For a typical GWAS, we have a phenotype of interest and we do not know any causal polymorphisms (loci) that affect this phenotype (but we would like to find them!)

- In an “ideal” GWAS experiment, we measure the phenotype and *N* genotypes THROUGHOUT the genome for *n* independent individuals

- To analyze a GWAS, we perform *N* independent hypothesis tests

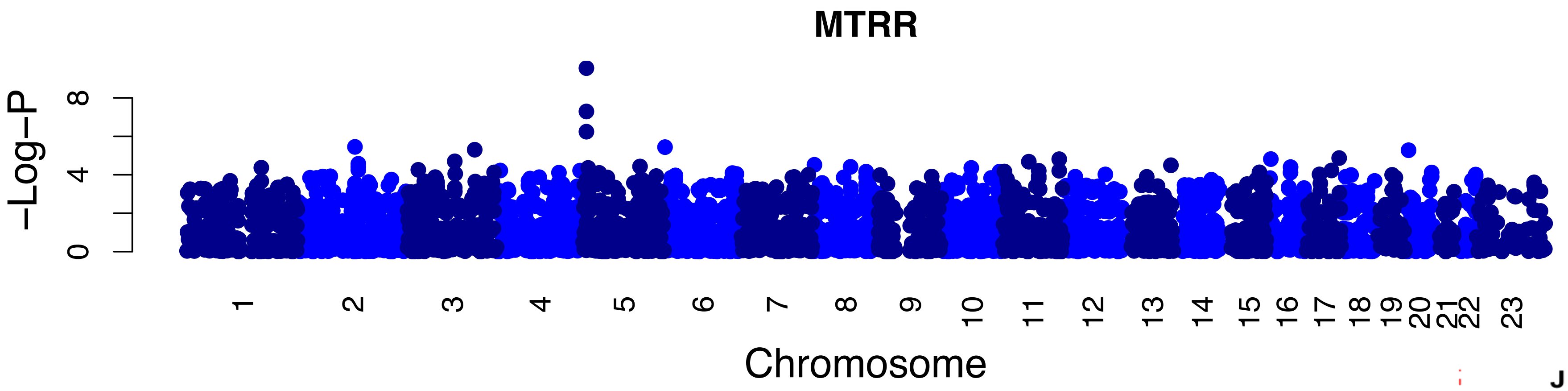
- When we reject the null hypothesis, we assume that we have located a position in the genome that contains a causal polymorphism (not the causal polymorphism!), hence a GWAS is a *mapping* experiment

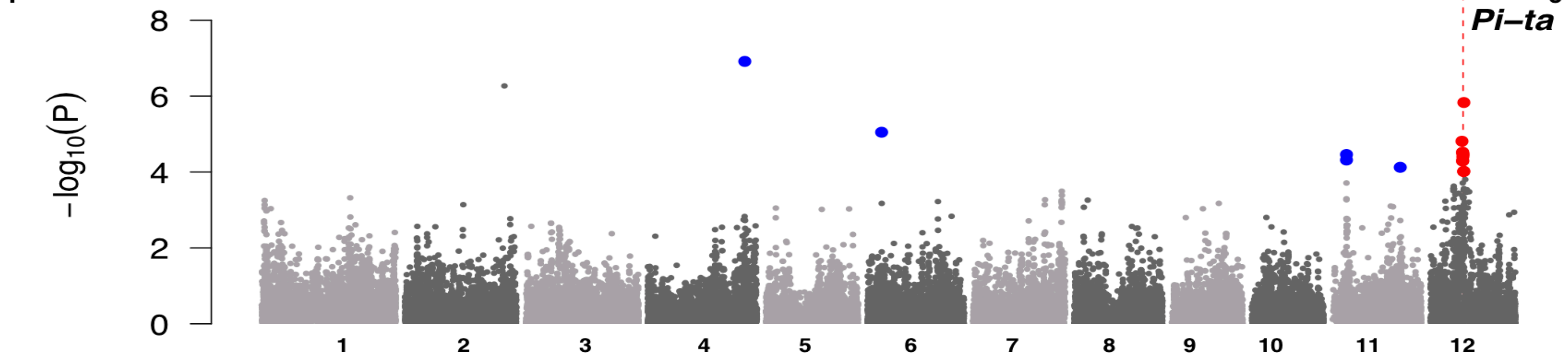
- This is as far as we can go with a GWAS (!!) such that (often) identifying the causal polymorphism requires additional data and or follow-up experiments, i.e. GWAS is a starting point

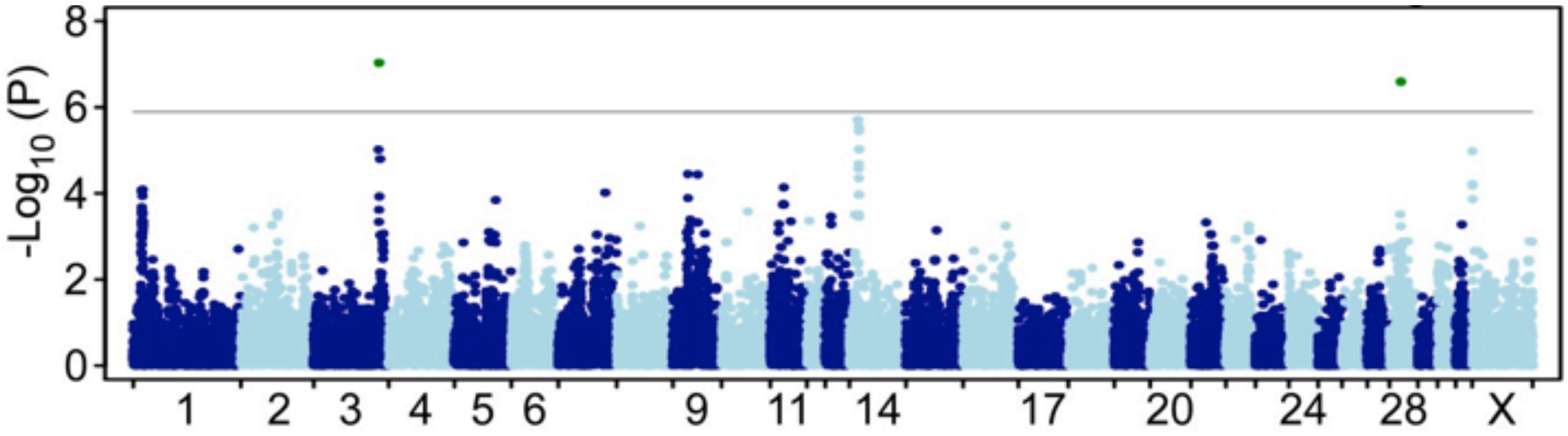
Interpreting “hits” from a GWAS analysis: measuring LD IV

- **Resolution** - the region of the genome indicated by significant tests for a set of correlated markers in a GWAS
- Recall that we often consider a set of contiguous significant markers (a “skyscraper” on a Manhattan plot) to indicate the location of a single causal polymorphism (although it need not indicate just one!)
- Note that the marker with the most significant p-value within a set is not necessarily closest to the causal polymorphism (!!)
- In practice, we often consider a set of markers with highly significant p-values to span the region where a causal polymorphism is located (but this is arbitrary and need not be the case!)
- In general, resolution in a GWAS is limited by the level of LD, which means there is a trade-off between resolution and the ability to map causal polymorphisms and that there is a theoretical limit to the resolution of a GWAS experiment (what is this limit?)

The Manhattan plot: examples

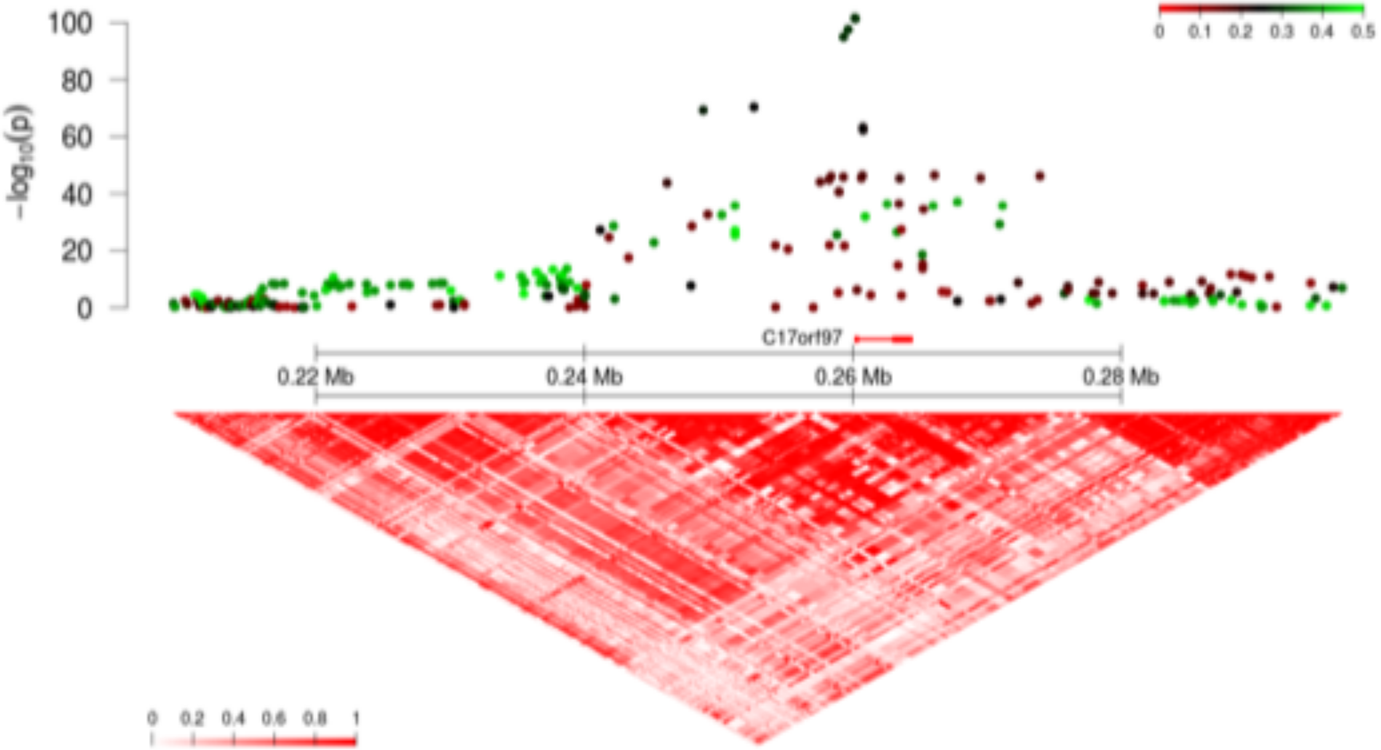






Patterns and representing LD

- 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!

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 (!!)

That’s it for today

- Next lecture (Thurs, March 23), we will continue our discussion of GWAS!