Quantitative Genomics and Genetics

BioCB 4830/6830; PBSB.5201.03

Lecture 18: GWAS analysis and statistical issues

Jason Mezey
March 26, 2024 (T) 8:40-9:55

Announcements

- There will be office hours tomorrow (Weds, March 27) 11AM-1PM
- Homework #4 (last homework!) is due by 11:59PM Friday March 29 (we will cover everything you need TODAY and THURSDAY!)
- Your midterm (!!) will be the week of April 8:
    - This is a take-home exam (!!) where you are allowed to use ANY sources available to you EXCEPT asking ANYONE about anything regarding the exam once it has started…
    - The midterm will be available after class on Tues, April 9 and will be due by noon on Thurs., April 11
    - What will be on the midterm? You will have to do a GWAS analysis JUST LIKE HOMEWORK #4 (!!), i.e., "SNPs" with two alleles such that genotypes are combinations of a, g, c, or t (e.g., cc, ct, tt, etc.), you'll have to code Xa and Xd, calculated MLE's and construct an F statistic for each of the N total SNPs, plot a Manhattan plot and a QQ plot (see next lecture!) AND interpret the data (btw NO COVARIATES = if you have the code to calculate an F statistic using Xa and Xd JUST LIKE HOMEWORK #4 you'll be good to go!)

Quantitative Genomics and Genetics - Spring 2024
BTRY 4830/6830; PBSB 5201.01

Midterm Exam

Available 11AM (ET), Tues., April 9
Due 11:59AM = 1 min before noon! (ET) Thurs., April 11

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 (even ChatGPT or similar!), 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 ED DISCUSSION!** (the only exceptions are Beulah, Sam, and Dr. Mezey, e.g., you MAY send us a private message on Canvas). 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.

2. Please pay attention to instructions and complete ALL requirements for ALL questions, e.g. some questions ask for R code, plots, AND written answers. We will give partial credit so it is to your advantage to attempt every part of every question.

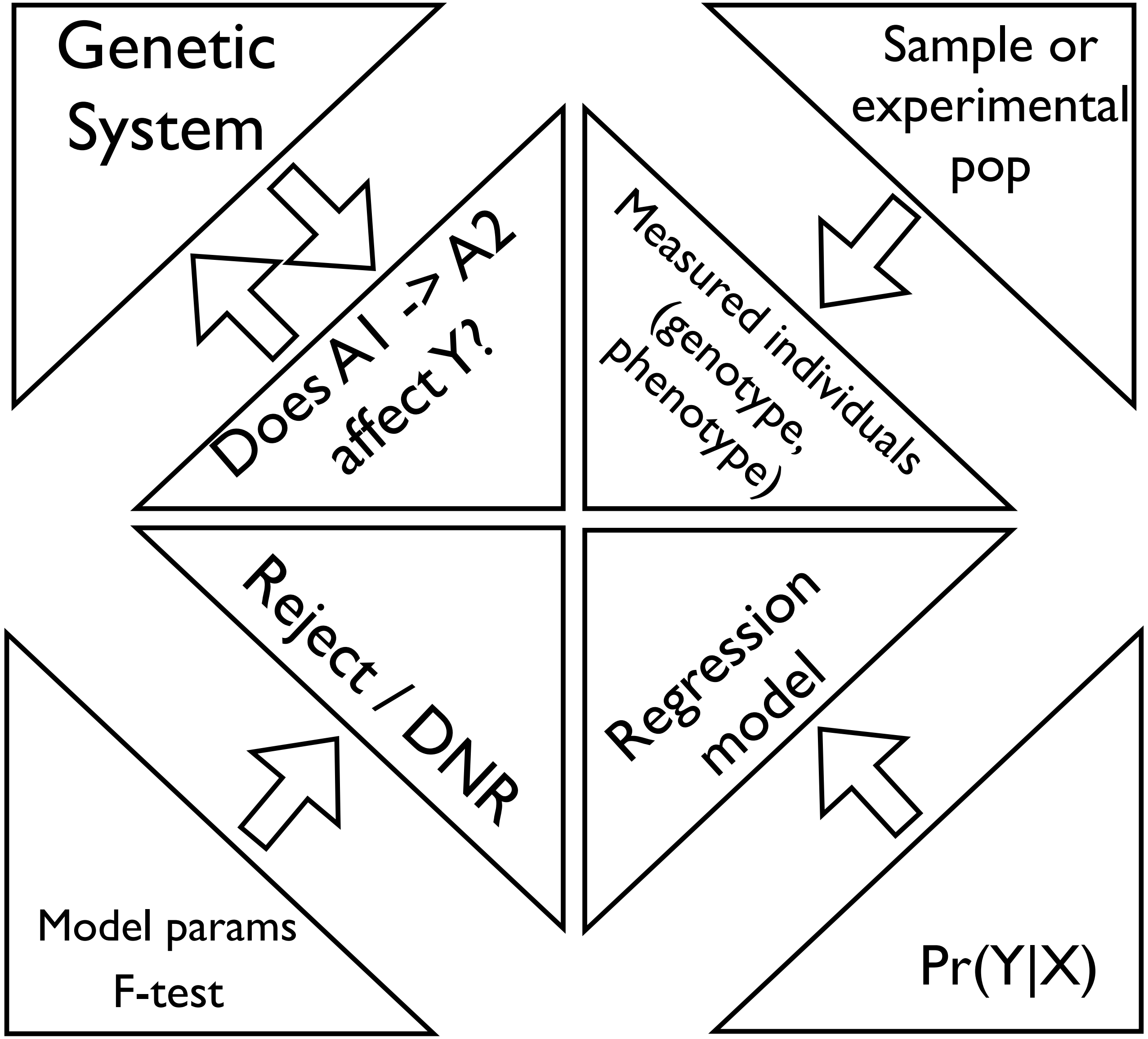
3. A complete answer to this exam will include R code answers, where you will submit your .Rmd script and the results of running your code in an associated .pdf file (plus an additional .pdf files if you have separate files for your written answers and code output). Note there will be penalties for scripts that fail to compile (!!). Also, as always, you do not need to repeat code for each part (i.e., if you write a single block of code that generates the answers for some or all of the parts, that is fine, but do please label your output that answers each question!!).

4. The exam must be uploaded on Canvas before 11:59AM (!!) = 1 minute before noon! (ET) Fri., March 31. It is your responsibility to make sure that it is in uploaded by then and no excuses will be accepted (power outages, computer problems, Cornell's internet slowed to a crawl, etc.). Remember: you are welcome to upload early! We will deduct points for being late for exams received after this deadline (even if it is by minutes!!).

Summary of lecture 18: GWAS Analysis

- Last lecture, we continued our introduction to Genome-Wide Association Studies (GWAS) where we discussed concepts of Linkage Disequilibrium (LD) and a Manhattan plot
- Today, we will also continue our introduction by discussing GWAS analysis issues, including those that relate to LD and statistical issues!

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: 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!)

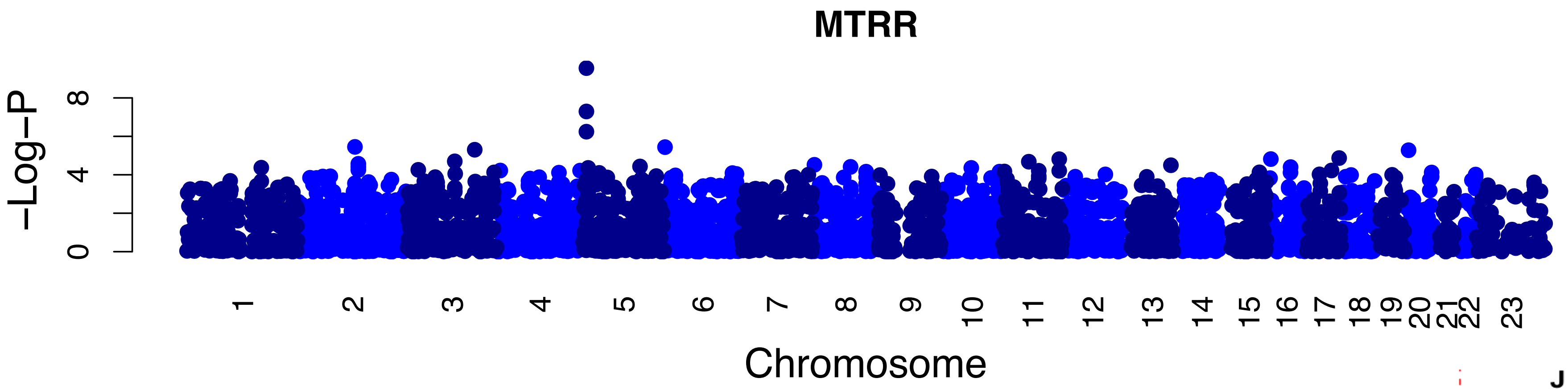
- In an “ideal” GWAS experiment, we measure the phenotype and *N* genotypes THROUGHOUT the genome for *n* independent individuals

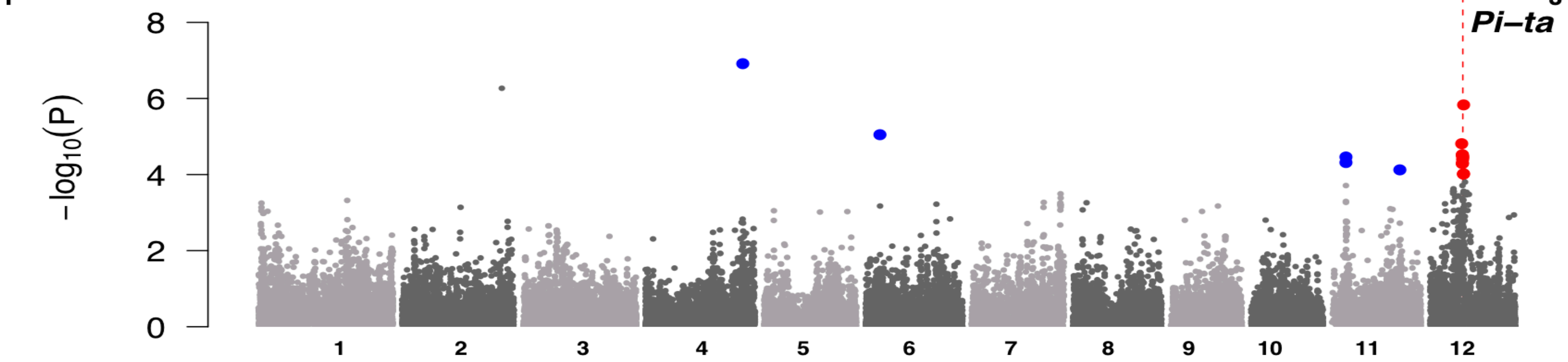
- To analyze a GWAS, we perform *N* independent hypothesis tests

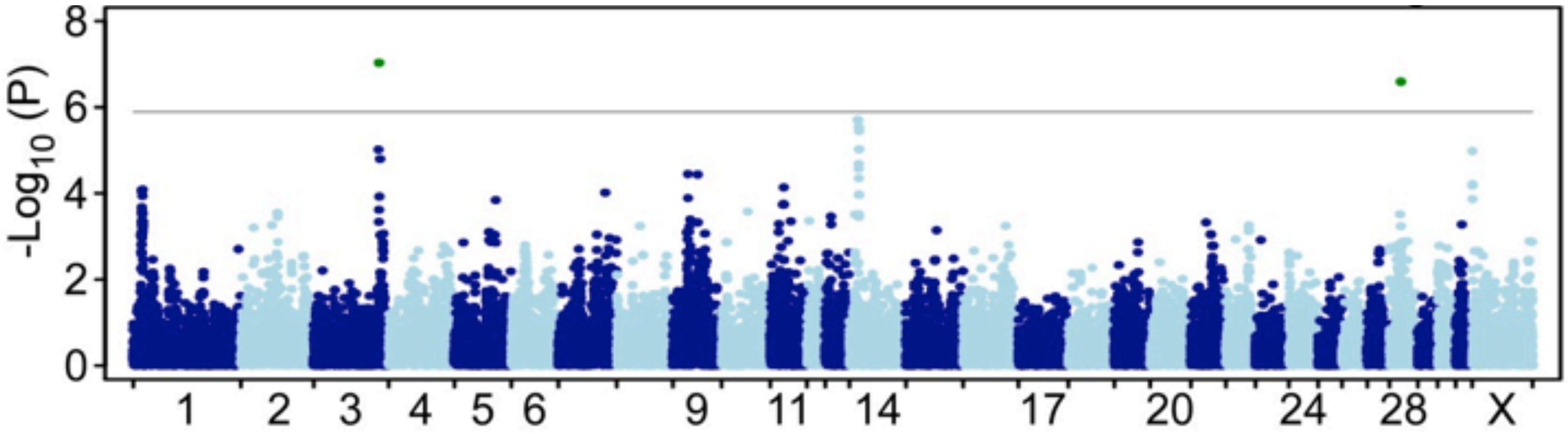
- When we reject the null hypothesis, we assume that we have located a position in the genome that contains a causal polymorphism (not the causal polymorphism!), hence a GWAS is a *mapping* experiment

- 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

Review: Manhattan plot



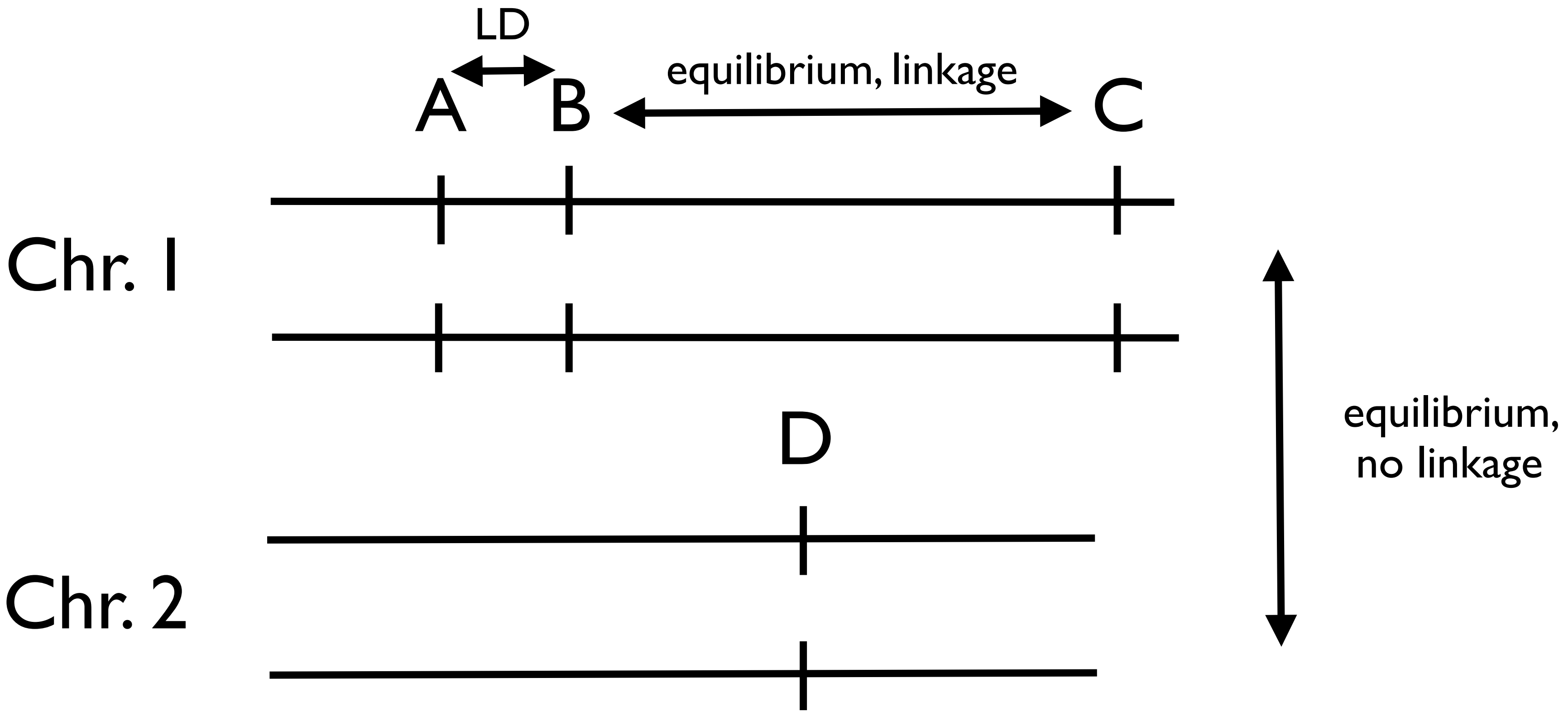




Review: Linkage Disequilibrium

- 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

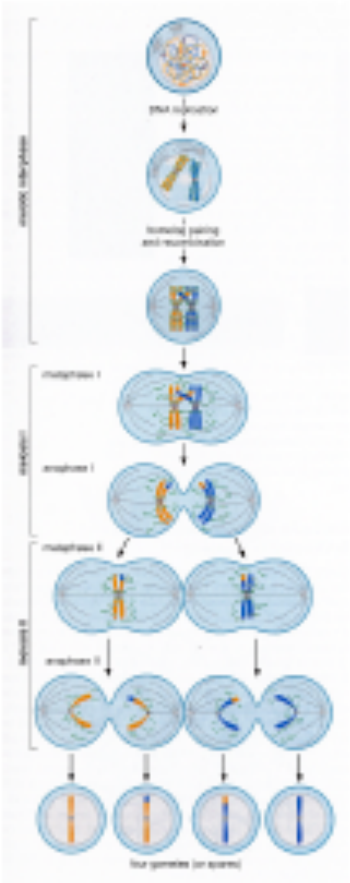


Population Genetic Causes of LD (in human populations)

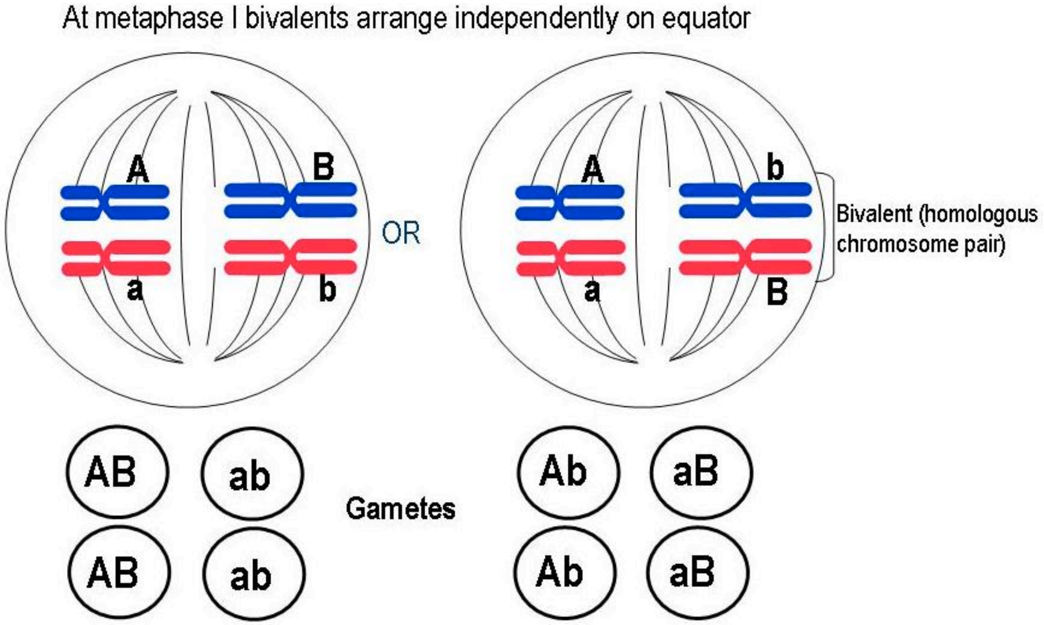
- The major factors responsible for patterns of LD in human populations are:
  - (1) Independent assortment of chromosomes
  - (2) “Random” mating
  - (3) Recombination
- Note (!!): this is the answer considering EXISTING variation in a population and therefore no MUTATION or MIGRATION
- Note that these 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



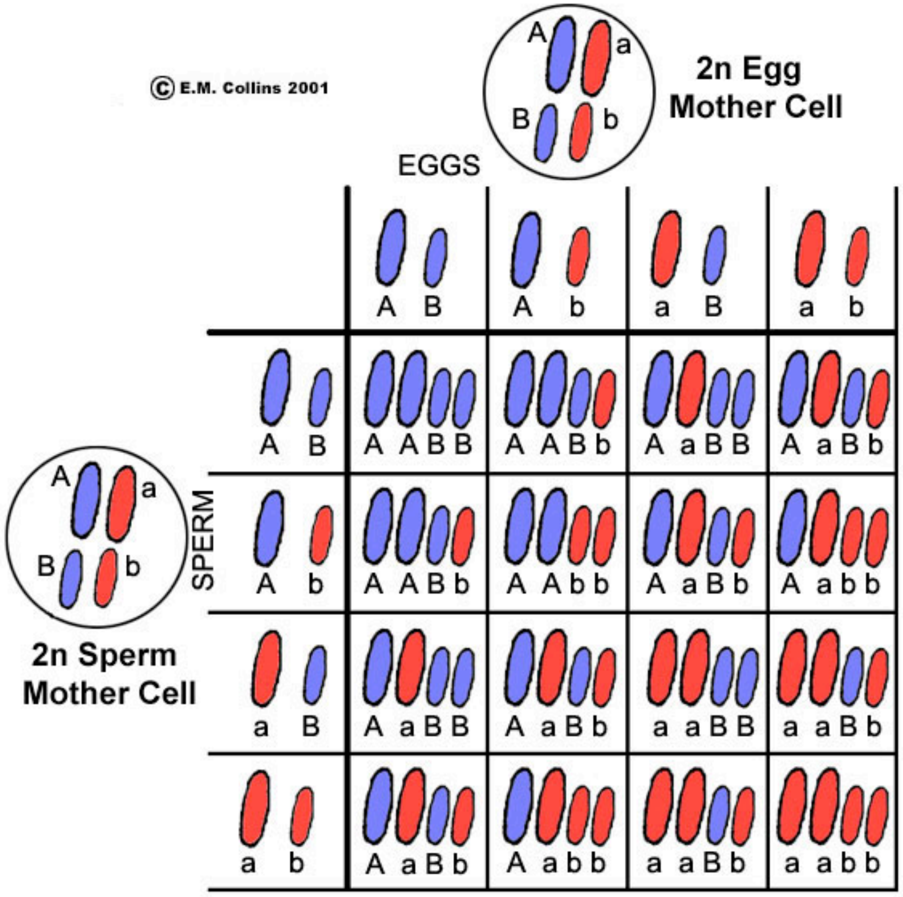
Credit: Watson et al., *Molecular Biology of the Gene*, CSHL Press, 2004


At metaphase I bivalents arrange independently on equator
A
a
B
b
OR
A
a
b
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:

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

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

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

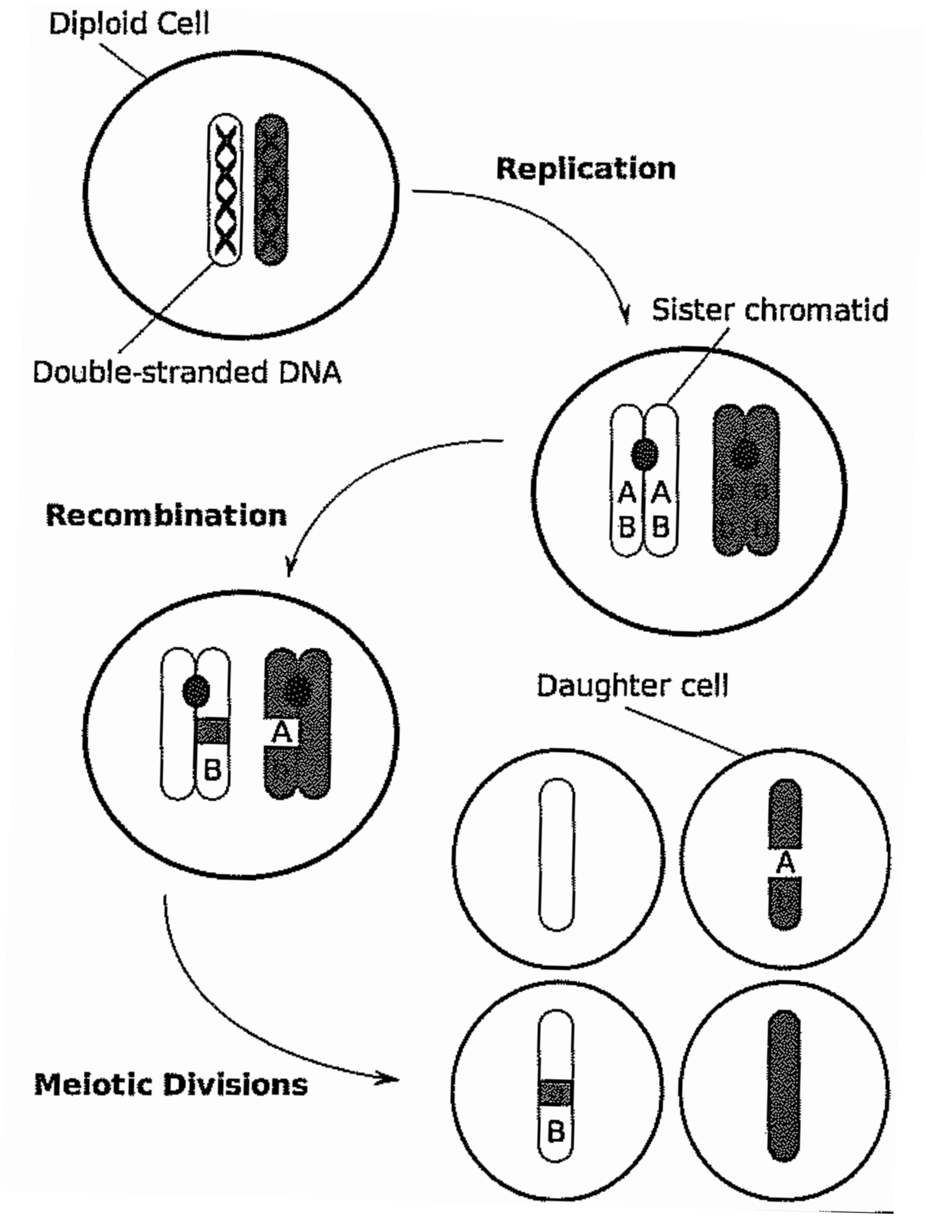
$$\cap (Corr(X_{d,A}, X_{a,B}) = 0) \cap (Corr(X_{d,A}, X_{d,B}) = 0)$$

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:

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

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

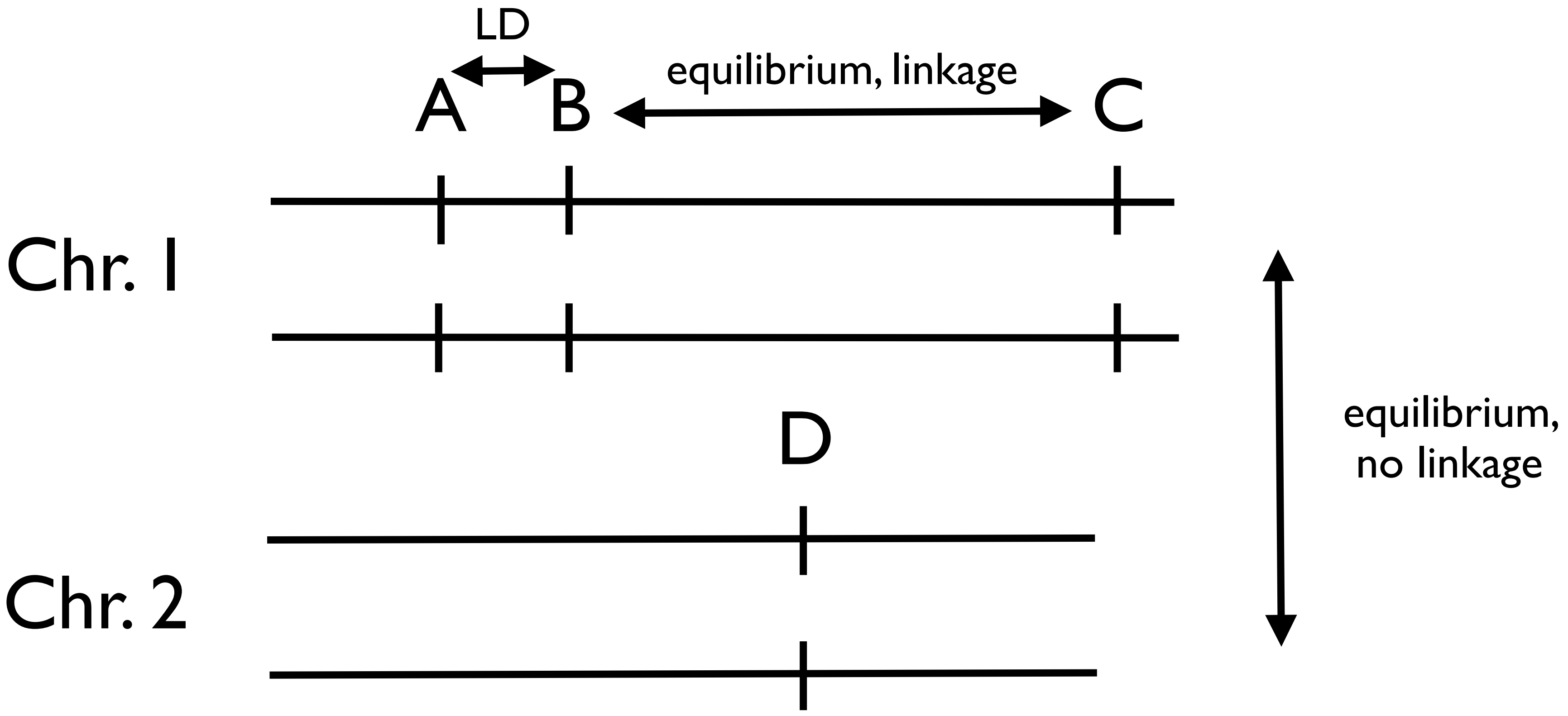
- Note recombination events disproportionally lower the probabilities of the more frequent pairs!

- 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):

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

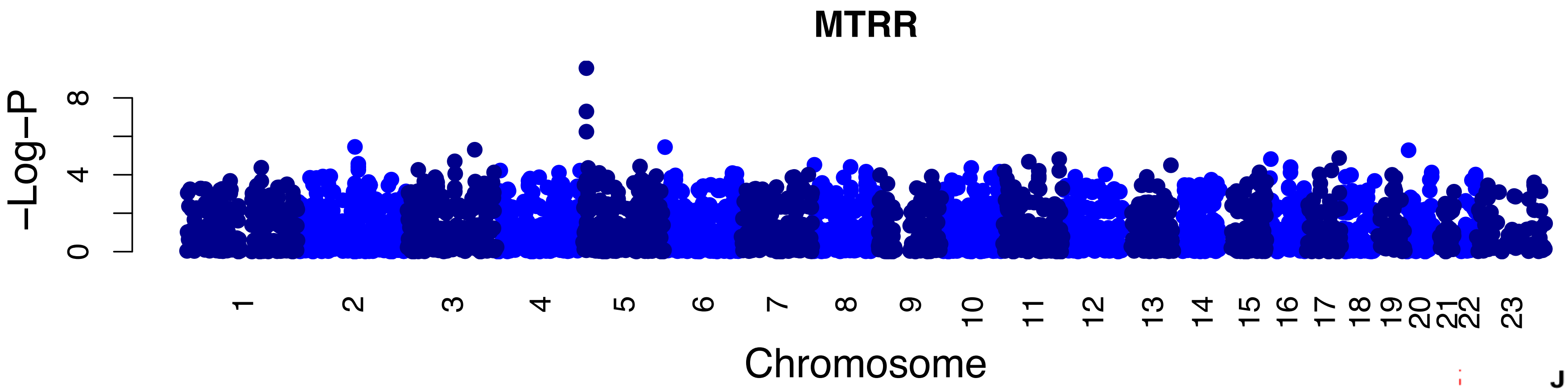
Reminder: Genome-Wide Association Study (GWAS)

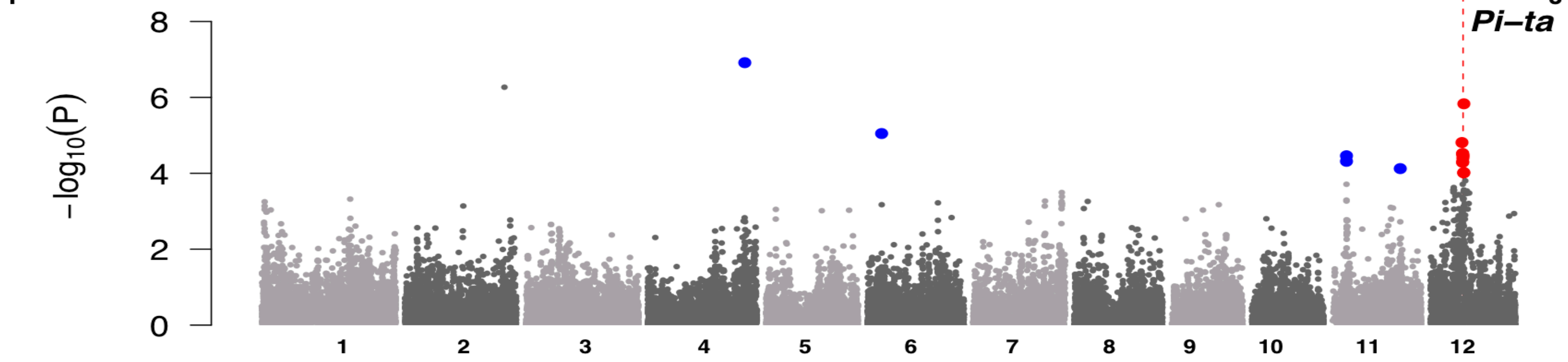
- For a typical GWAS, we have a phenotype of interest and we do not know any causal polymorphisms (loci) that affect this phenotype (but we would like to find them!)
- In an "ideal" GWAS experiment, we measure the phenotype and *N* genotypes THROUGHOUT the genome for *n* independent individuals
- To 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

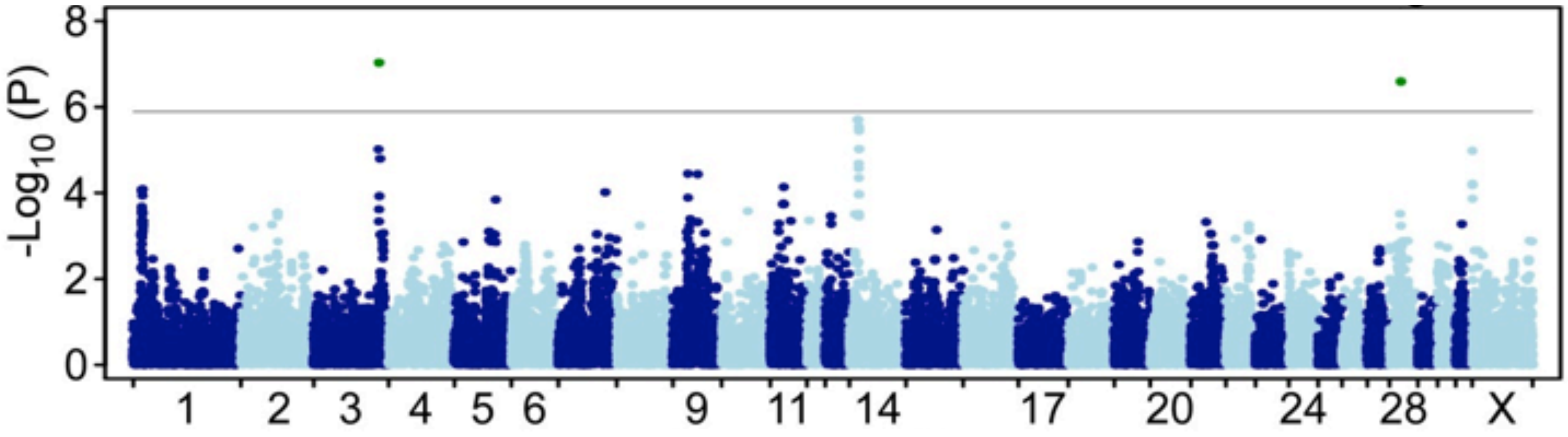
Interpreting "hits" from a GWAS analysis: measuring LD IV

- **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



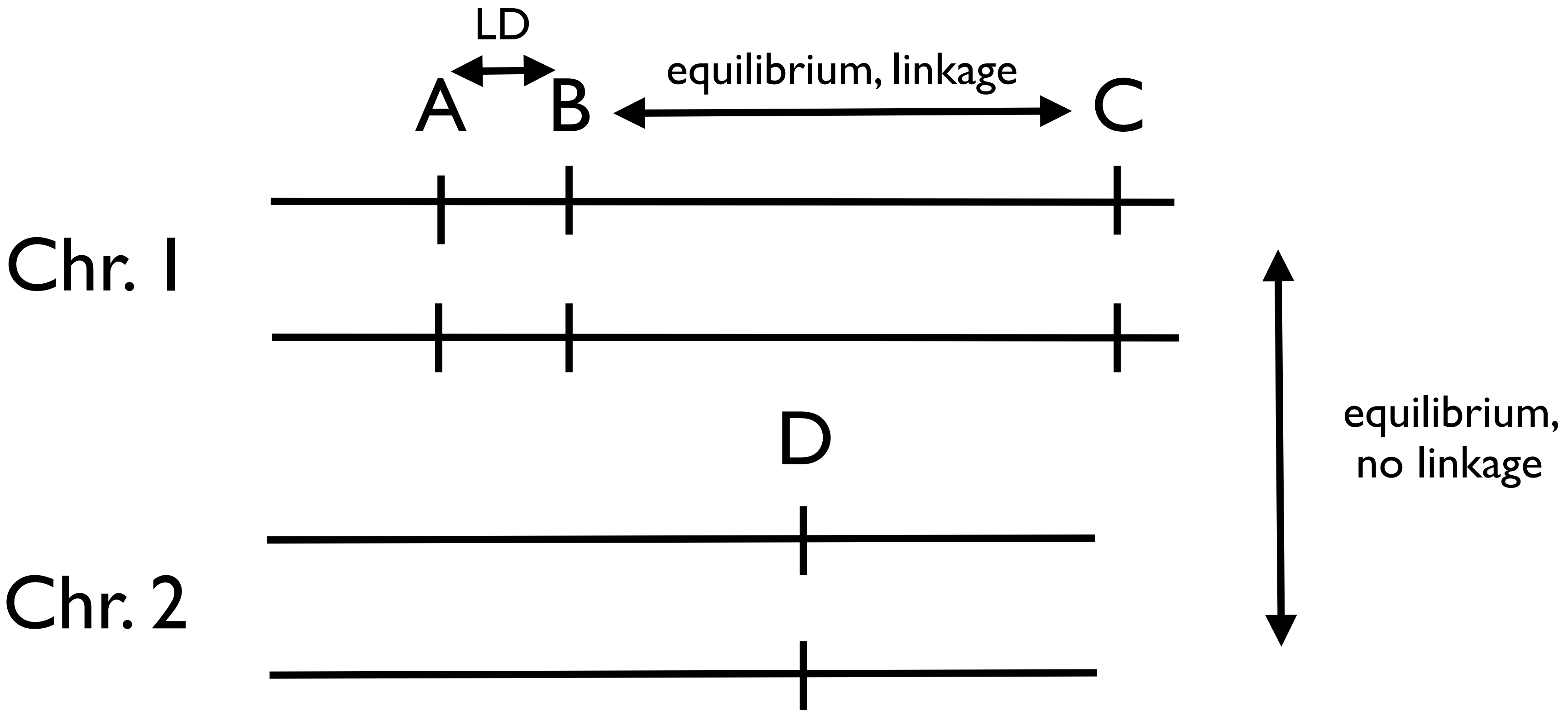




Linkage Disequilibrium

- 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



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:

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

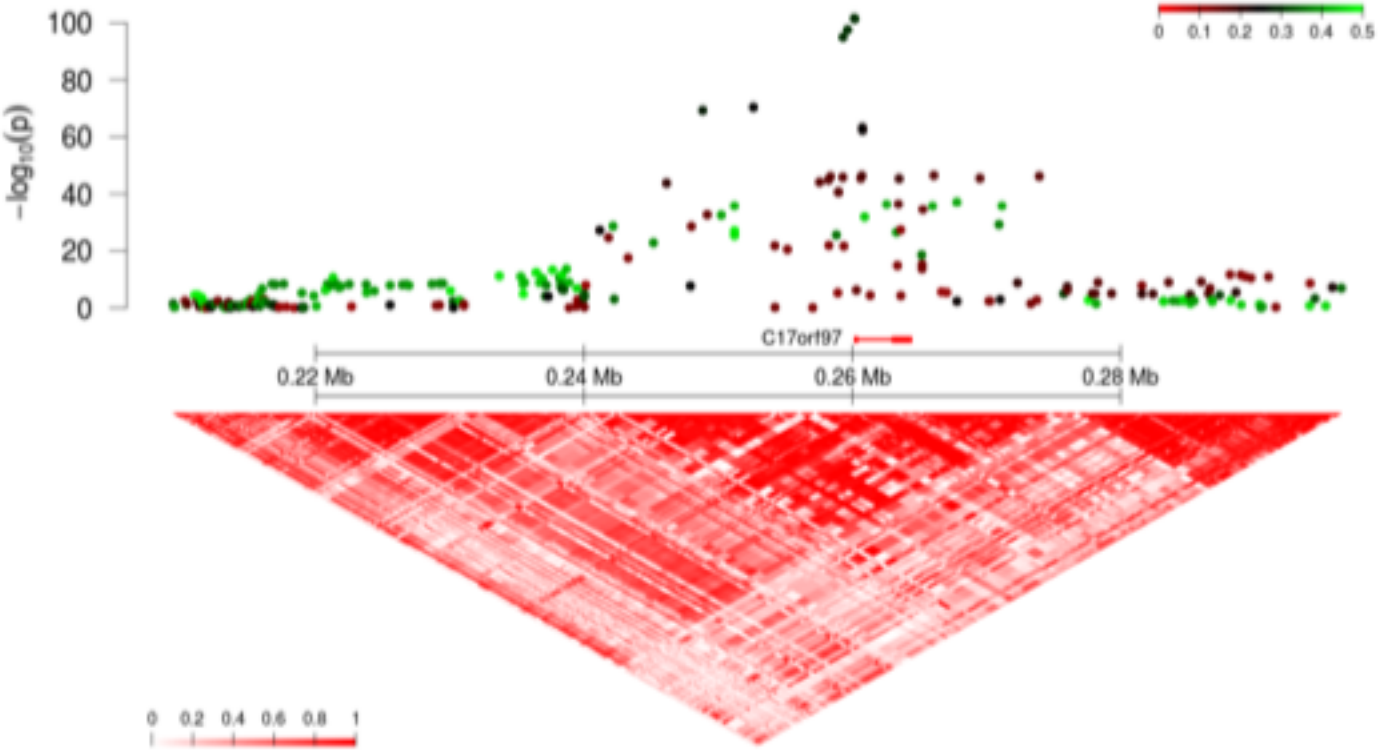
$$D' = \frac{D}{min(Pr(A_1 B_2), Pr(A_2, B_1))} \text{if} D > 0$$

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

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

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:



Issues for successful mapping of causal polymorphisms in GWAS

- For GWAS, we are generally concerned with correctly identifying the position of as many causal polymorphisms as possible (True Positives) while minimizing the number of cases where we identify a position where we think there is a causal polymorphism but there is not (False Positive)
- We are less concerned with cases where there is a causal polymorphism but we do not detect it (why is this?)
- Issues that affect the number of True Positives and False Positives that we identify in a GWAS can be statistical and experimental (or a combination)

Statistical Issues 1: Type 1 error

- Recall that Type 1 error is the probability of incorrectly rejecting the null hypothesis when it is correct
- A Type 1 error in a GWAS produces a false positive
- We can control Type 1 error by setting it to a specified level but recall there is a trade-off: if we set it to low, we will not make a Type 1 error but we will also never reject the null hypothesis, even when it is wrong (e.g. if Type 1 error is to low, we will not detect ANY causal polymorphisms)
- In general we like to set a conservative Type 1 error for a GWAS (why is this!?)
- To do this, we have to deal with the *multiple testing problem*

Statistical Issues II: Multiple Testing

- Recall that when we perform a GWAS, we perform *N* hypothesis tests (where *N* is the number of measured genotype markers)
- Also recall that if we set a Type 1 error to a level (say 0.05) this is the probability of incorrectly rejecting the null hypothesis
- If we performed *N* tests that were independent, we would therefore expect to incorrectly reject the null *N**0.05 and if N is large, we would therefore make LOTS of errors (!!)
- This is the multiple testing problem = the more tests we perform the greater the probability of making a Type 1 error
- Now in a GWAS, our tests are not independent (LD!) but we could still make many errors by performing N tests if we do not set the Type I error appropriately

Correcting for multiple tests I

- Since we can control the Type I error, we can correct for the probability of making a Type 1 error due to multiple tests
- There are two general approaches for doing this in a GWAS: those that involve a *Bonferroni correction* and those that involve a correction based on the estimate the *False Discovery Rate* (FDR)
- Both are different techniques for controlling Type 1 error but in practice, both set the Type I error to a specified level (!!)

Correcting for multiple tests II

- A Bonferroni correction sets the Type I error for the entire GWAS using the following approach: for a desired type 1 error α set the Bonferroni Type 1 error α_B to the following:

$$\alpha_B = \frac{\alpha}{N}$$

- We therefore use the Bonferroni Type I error to assess EACH of our *N* tests in a GWAS

- For example, if we have *N*=100 in our GWAS and we want an overall GWAS Type 1 error of 0.05, we require a test to have a p-value less than 0.0005 to be considered significant

That’s it for today

- Next lecture, we will continue our discussion of issues in GWAS analysis (!!)