Quantitative Genomics and Genetics

BioCB 4830/6830; PBSB.5201.03

Lecture 17: GWAS analysis

Jason Mezey
March 21, 2024 (Th) 8:40-9:55

Announcements

- REMINDER: I will be lecturing by zoom next week (March 26 and 28)
- There will be office hours next week as scheduled
- Homework #4 (last homework!) is due by 11:59PM Friday March 29
- 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
  - I will tell you more about what will be on the midterm today in following lectures…

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.

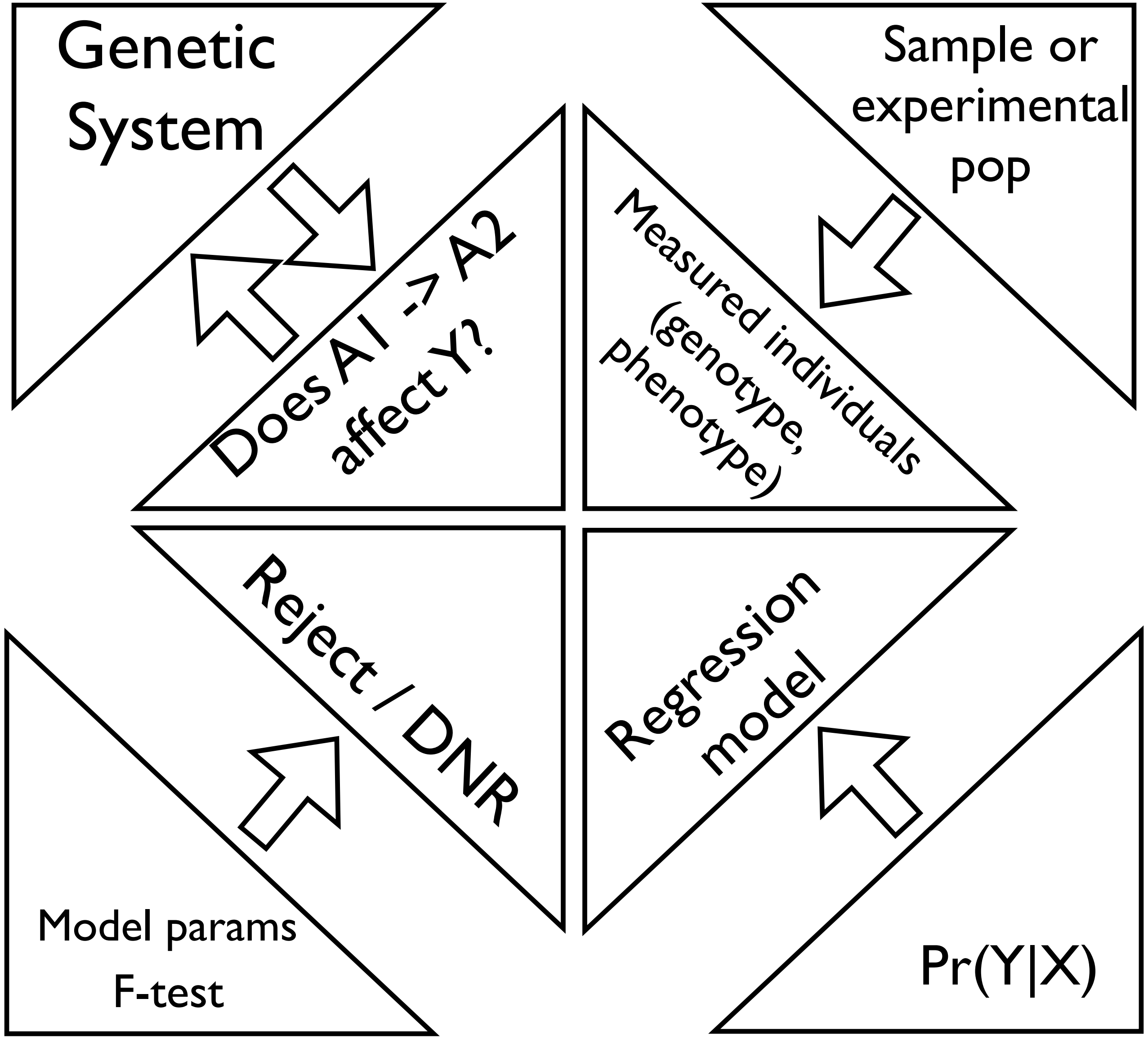
3. A complete answer to this exam will include R code answers, where you will submit your .Rmd script and the results of running your code in an associated .pdf file (plus an additional .pdf files if you have separate files for your written answers and code output). Note there will be penalties for scripts 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) Thurs, April 11. 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 17: GWAS Analysis

- Last lecture, we began our introduction to Genome-Wide Association Studies (GWAS)!
- Today, we will continue the discussion including introduction to the concepts of Linkage Disequilibrium (LD) and a Manhattan plot
- We will also begin our introduction to statistical and other issues important for 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)

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: Quantitative genomic analysis I

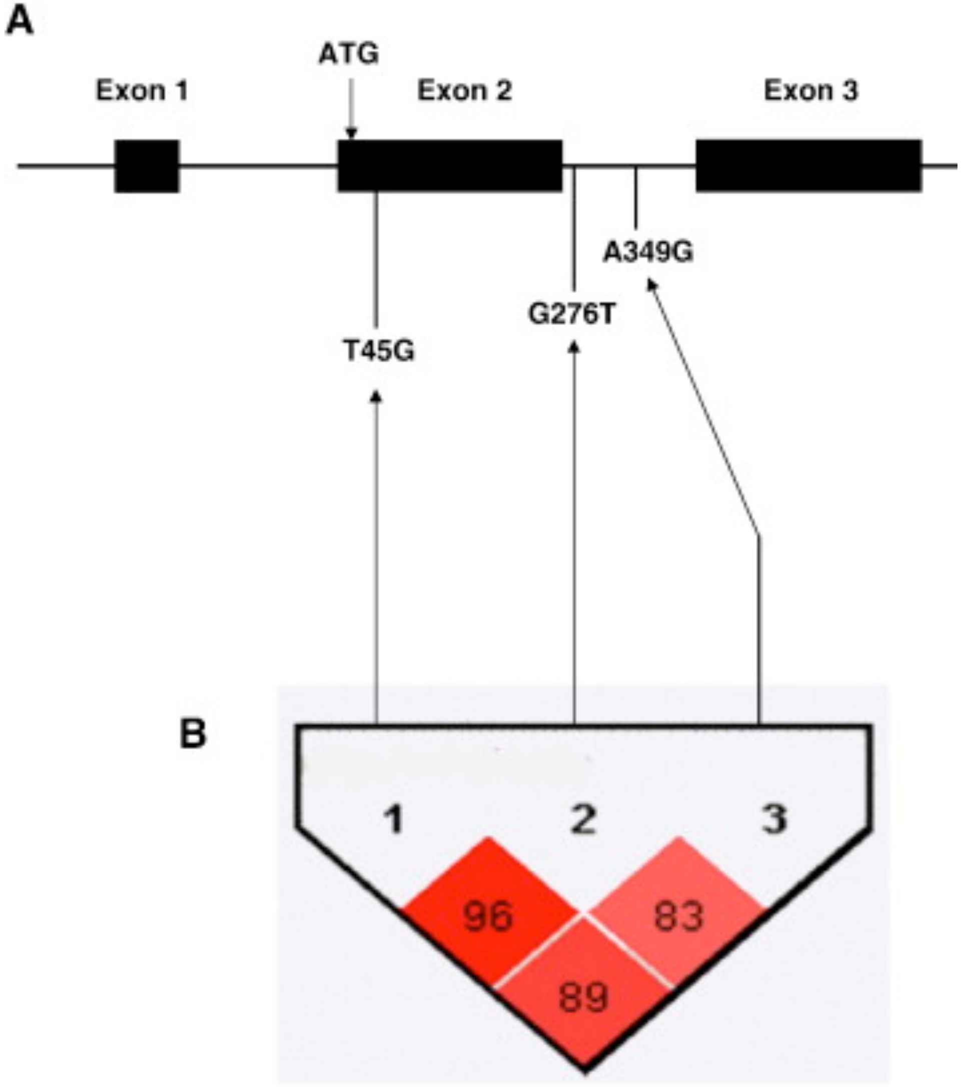
- We now know how to assess the null hypothesis as to whether a polymorphism has a causal effect on our phenotype
- Occasionally we will assess this hypothesis for a single genotype
- In quantitative genomics, we generally do not know the location of causal polymorphisms in the genome
- We therefore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

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

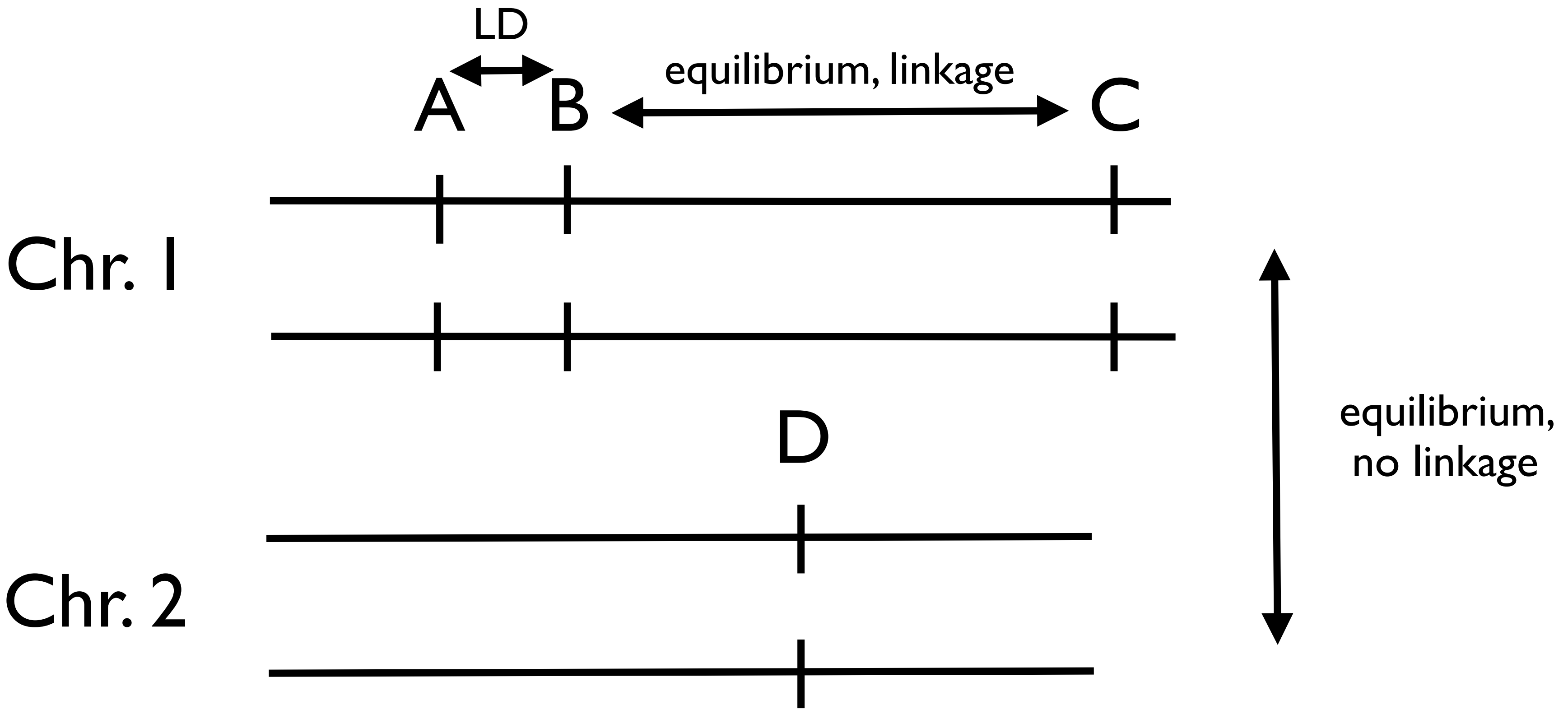
- If we test a (non-causal) genotype that is correlated with the causal genotype AND if correlated genotypes are in the same position in the genome THEN we can identify the genomic position of the casual genotype (!!)
- This is the case in genetic systems (why!?)
- Do we know which genotype is causal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



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!)

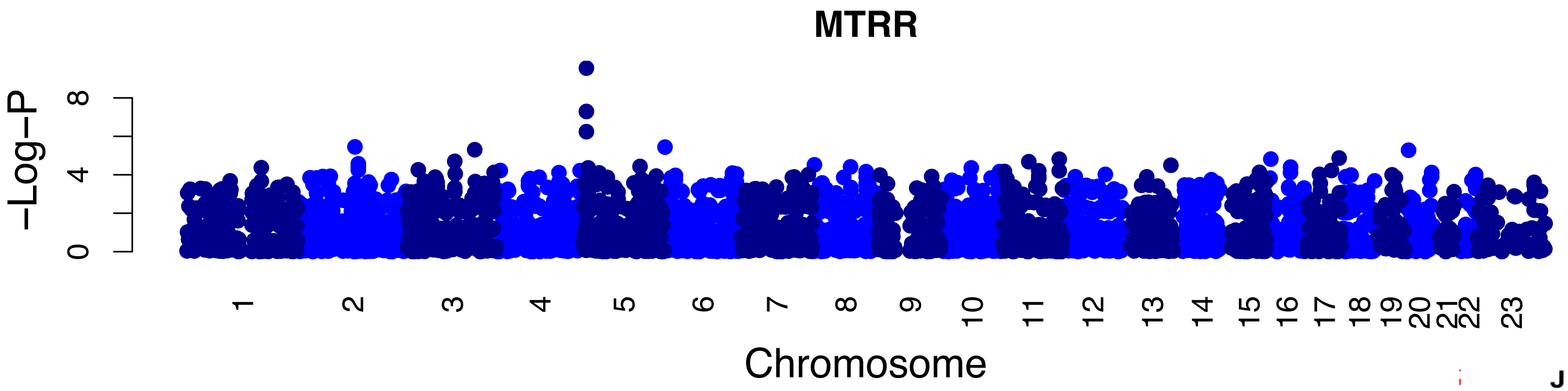
- In an "ideal" GWAS experiment, we measure the phenotype and *N* genotypes THROUGHOUT the genome for *n* independent individuals

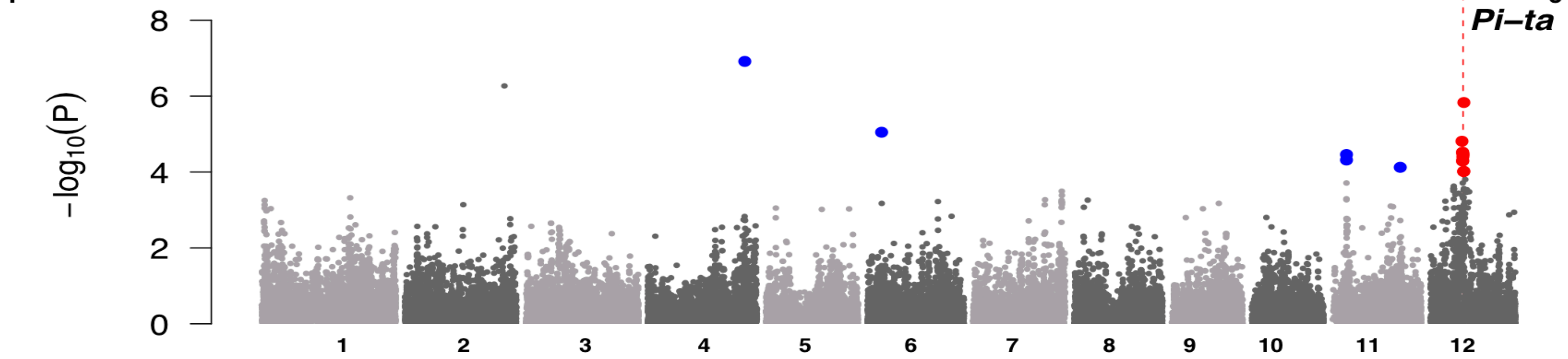
- To analyze a GWAS, we perform *N* independent hypothesis tests

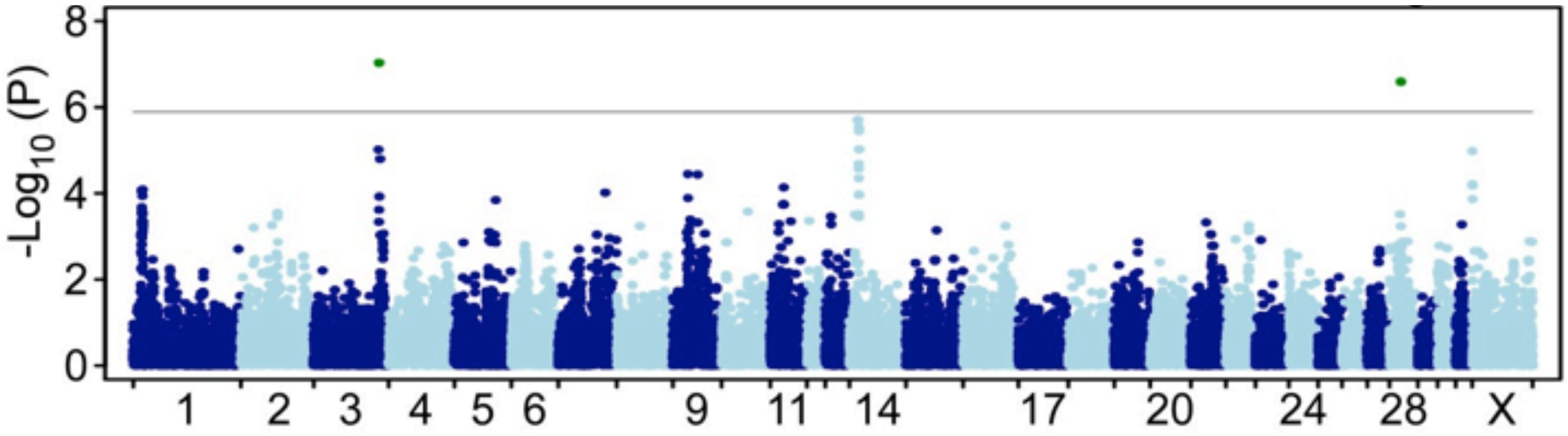
- When we reject the null hypothesis, we assume that we have located a position in the genome that contains a causal polymorphism (not the causal polymorphism!), hence a GWAS is a *mapping* experiment

- 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

The Manhattan plot: examples



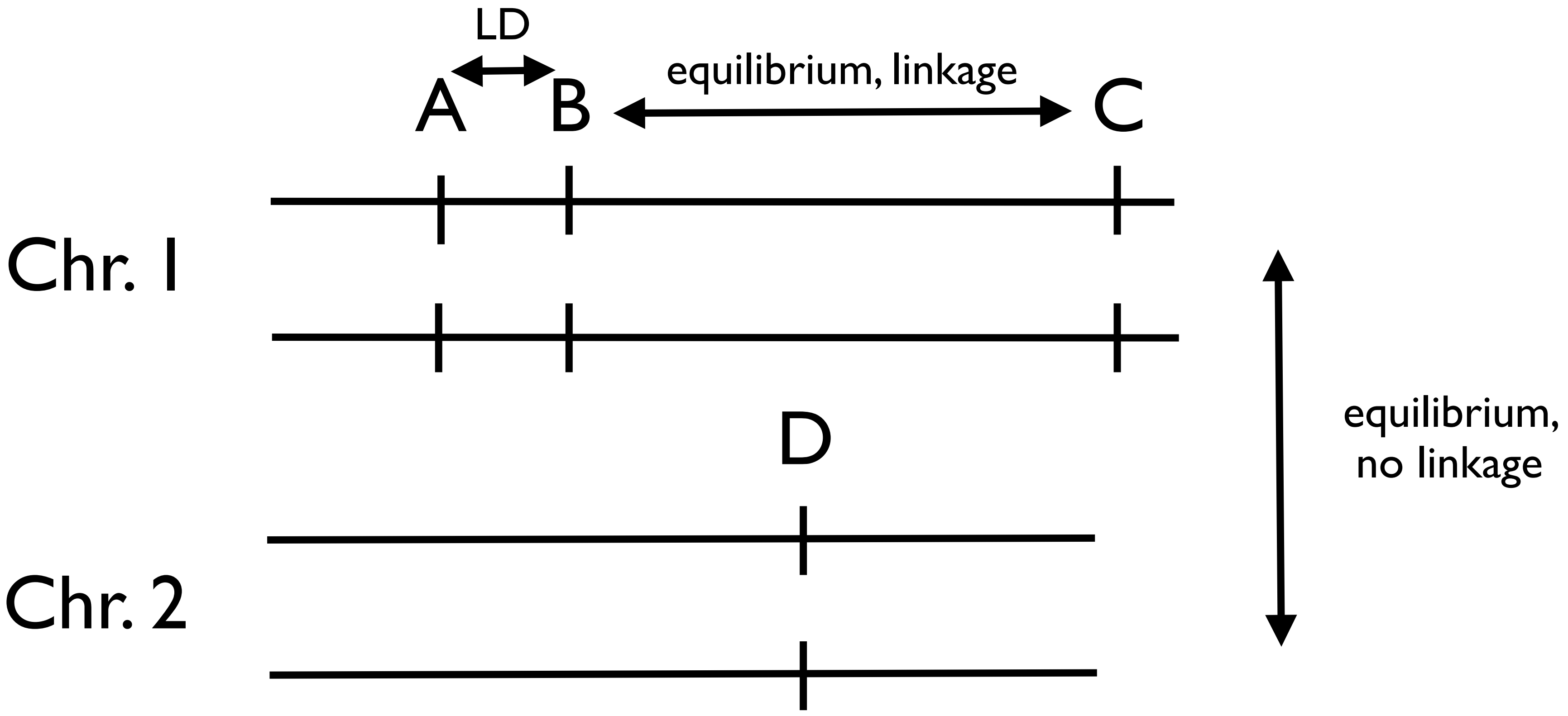




Linkage Disequilibrium

- 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



That’s it for today

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