Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.52201.03

Lecture 17: Introduction to GWAS II

Jason Mezey

March 23, 2023 (Th) 8:05-9:20

Announcements

- Homework related announcements (!!)
  - Homework #4 is due 11:59pm (Mon) March 27
  - We will cover the last topics you need for QUESTION #2 today (!!)
  - PLEASE run your code and SUBMIT a pdf of the results of running your code (as well as your rmd), i.e., if you submit a compiled latex pdf with your "written" answers please submit another pdf with your code output (you can submit two pdfs on CMS!)
  - A key for homework #4 will be posted on Tues (March 28) (!!)
  - As a consequence, penalties for LATE homework #4 submissions will be considerable (so PLEASE get your homework in on time!)
  - The midterm (!!) will ask you to do similar tasks as needed for homework #4 (so it is to your advantage to spend time on it!)
  - For computer labs today Thurs (March 23) and tomorrow (March 24) will review and cover what you need to complete your homework (and what you will need for the midterm!)
  - I will hold office hours this coming Mon (March 27) 12:30-2:30 where I am happy to discuss homework #4 or any other topics

Announcements

- Midterm Exam related announcements (!!)
  - 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!
  - We WILL have lecture this coming Tues (March 28) but we will NOT have lecture Thurs (March 30) and we will NOT have computer labs next week (and we will not have lectures or labs the following week = Cornell, Ithaca Spring break)
  - See next slides for more information on the midterm…

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 <u>YOU ARE NOT ALLOWED TO</u> COMMUNICATE OR IN ANY <u>WAY ASK ANYONE FOR ASSISTANCE WITH</u> 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.

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.

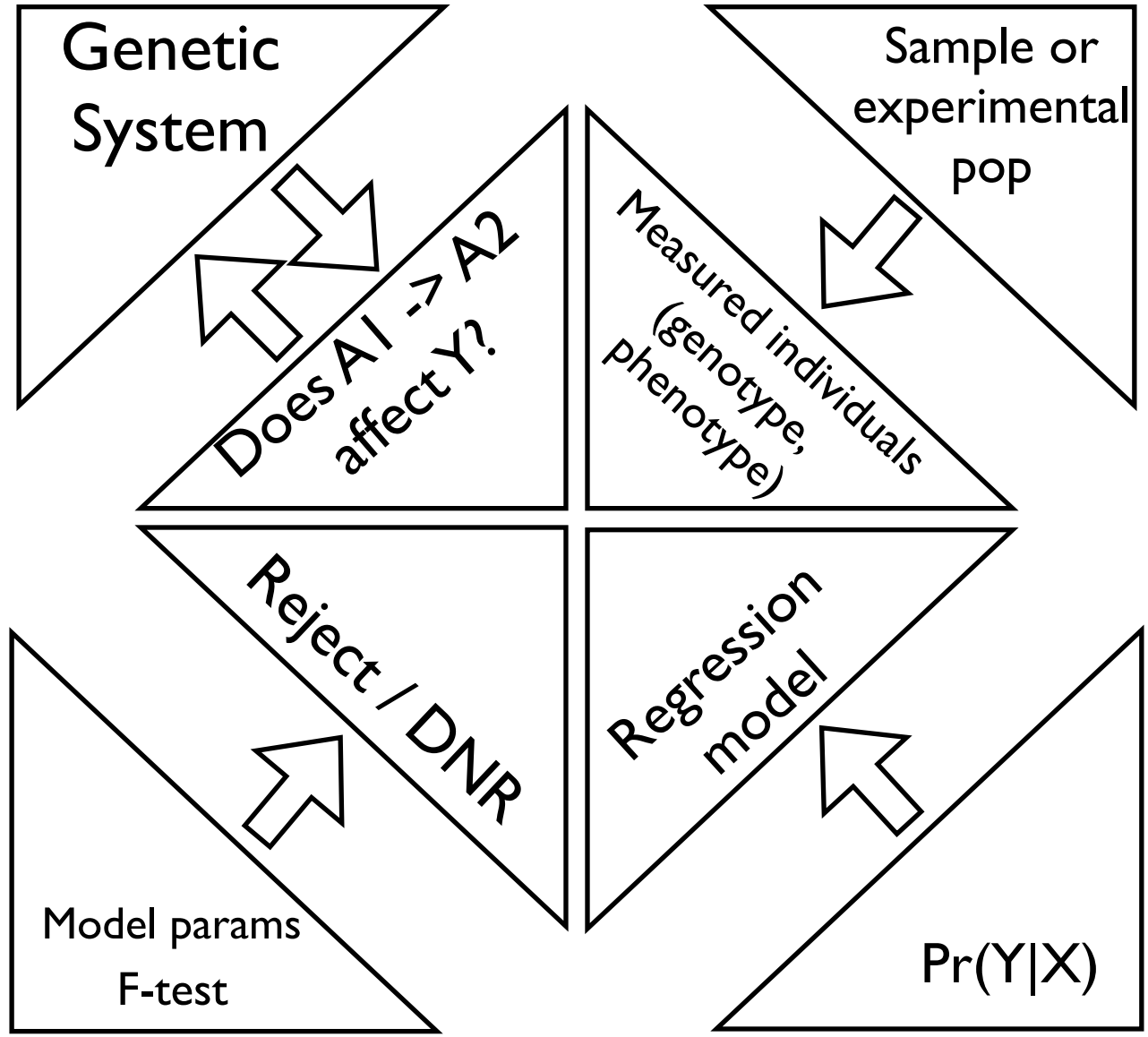
3. A complete answer to this exam will include R code answers, where you will submit your .Rmd script and the results of running your code in an associated .pdf file (plus an additional .pdf files if you have separate files for your written answers and code output). Note there will be penalties for scripts 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 CMS before 11:59PM (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 17: Introduction to GWAS

- Last lecture, we began our discussion of GWAS!
- Today, we will continue with our discussion of GWAS by discussing statistical and experimental issues impacting the success of a 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:

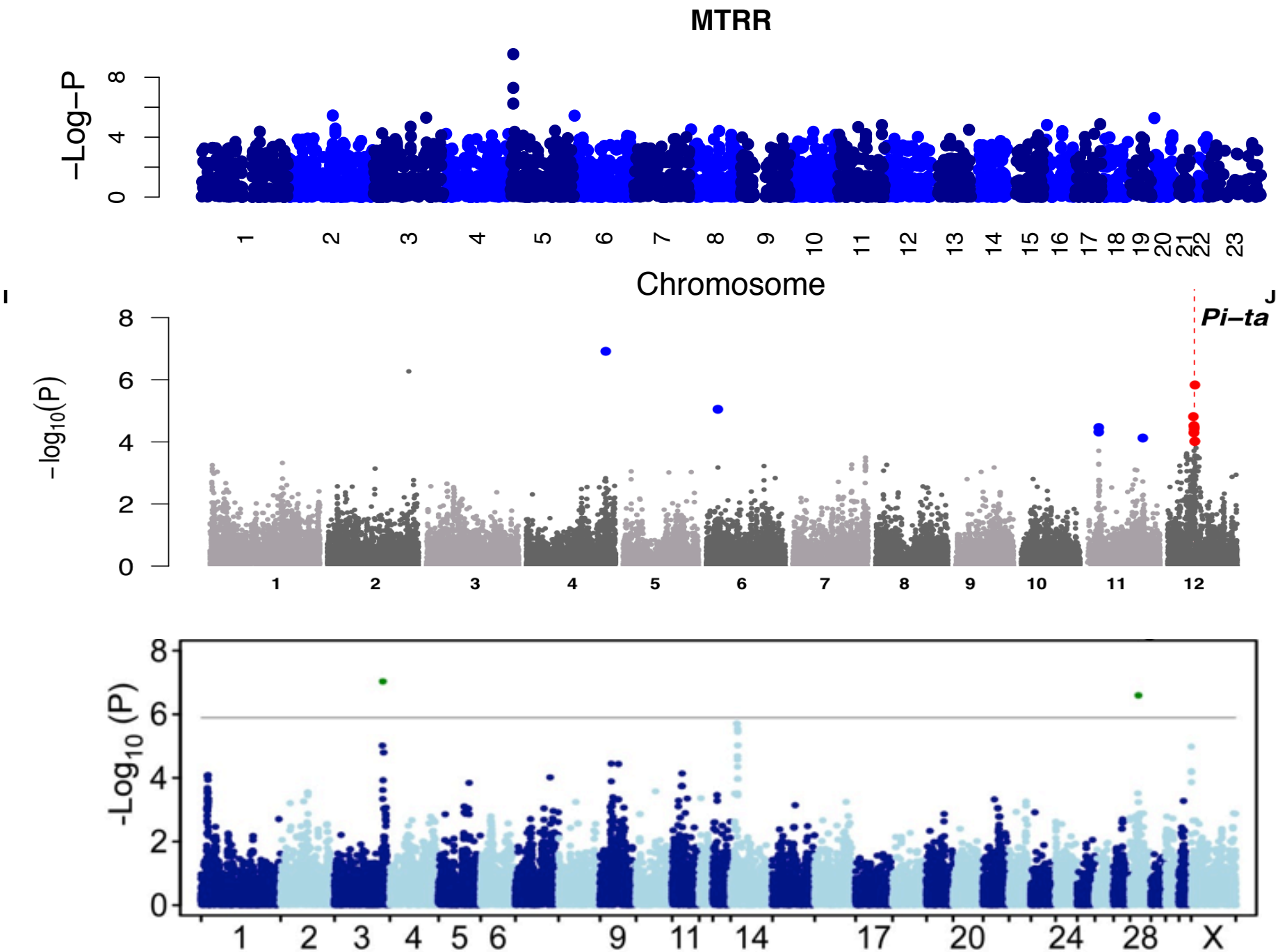
$$A_1 \rightarrow A_2 \Rightarrow \Delta Y | Z$$

- 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: Quantitative genomic analysis I

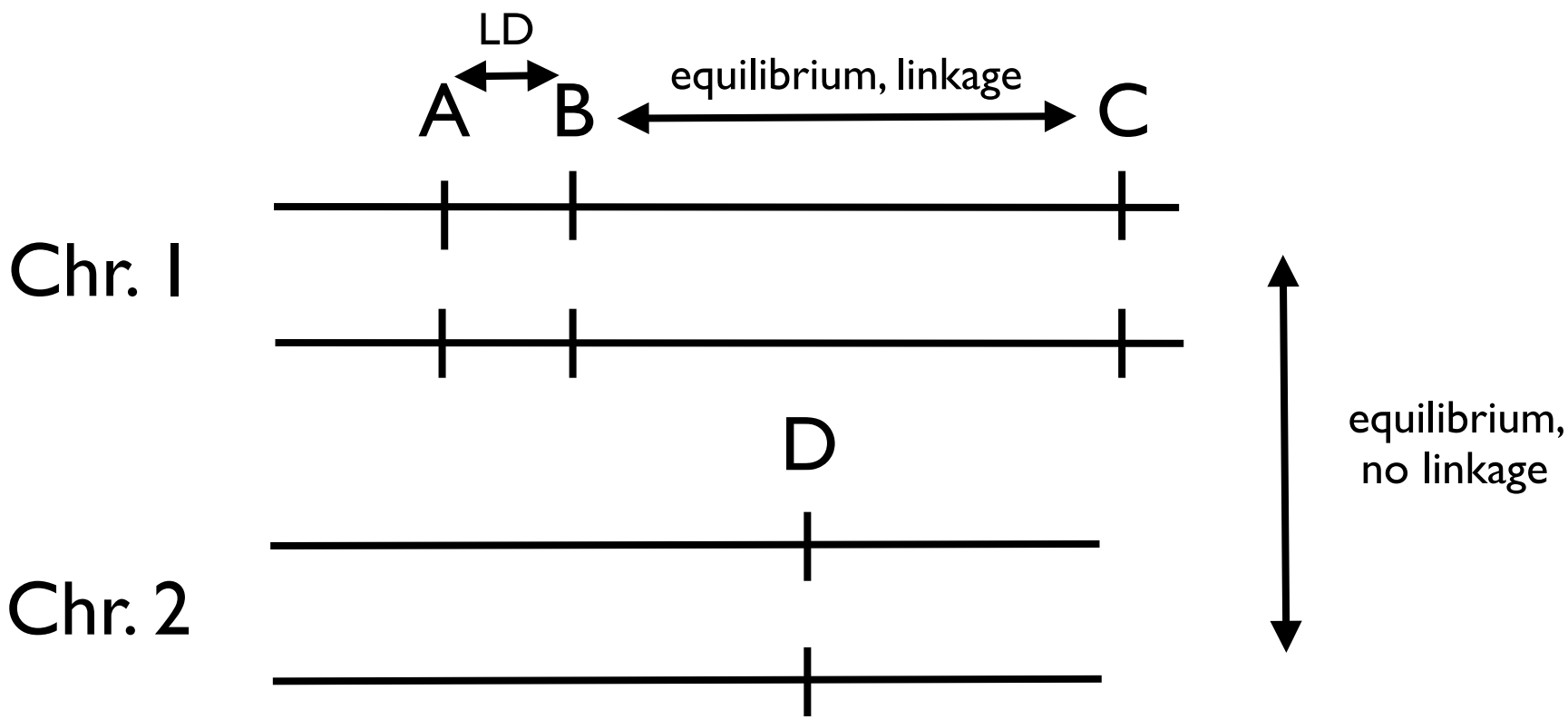
- We now know how to assess the null hypothesis as to whether a polymorphism has a causal effect on our phenotype
- Occasionally we will assess this hypothesis for a single genotype
- In quantitative genomics, we generally do not know the location of causal polymorphisms in the genome
- We therefore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

Review: 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



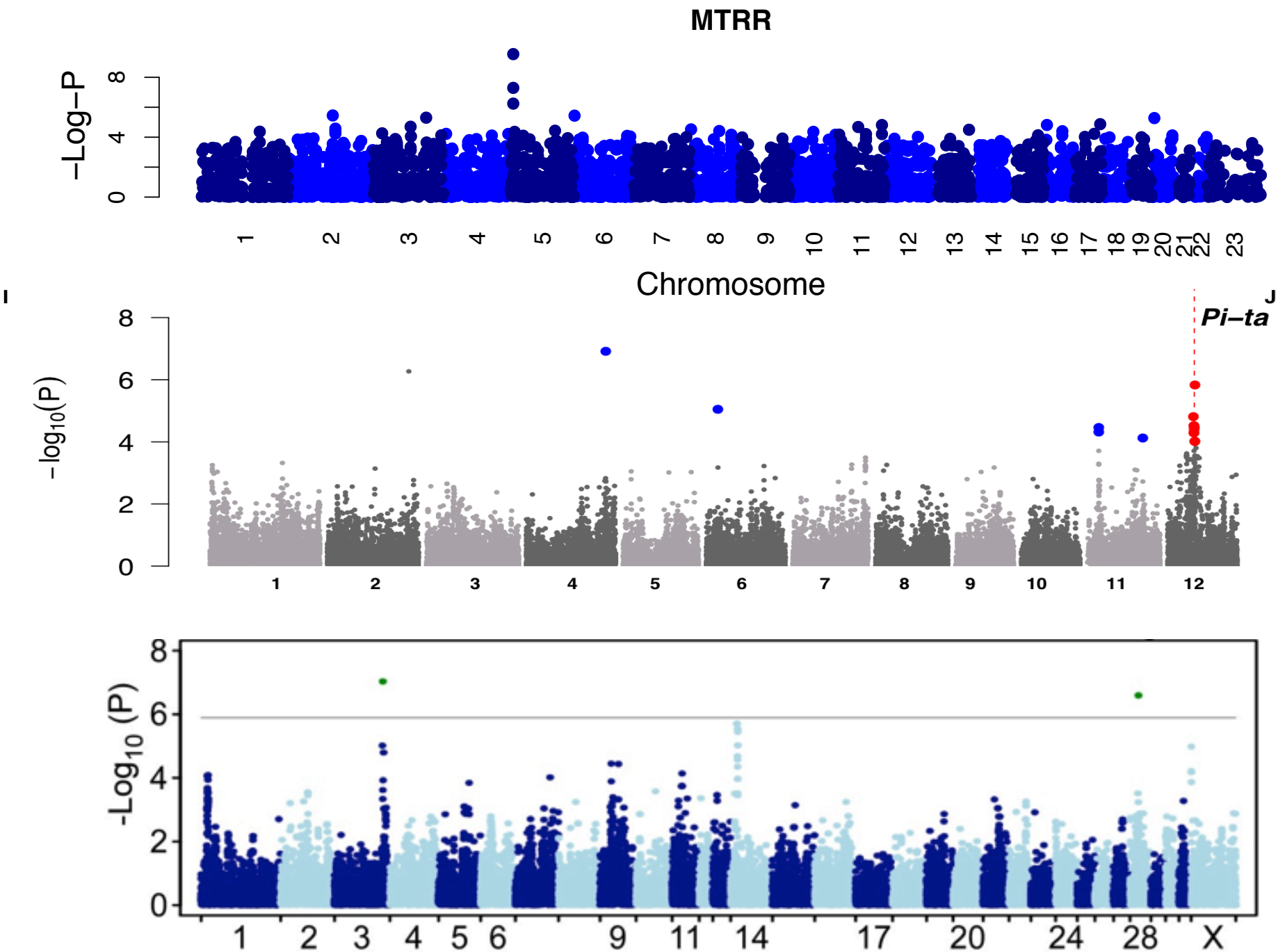
Review: Genome-Wide Association Study (GWAS)

- 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: Interpreting "hits" from a GWAS analysis

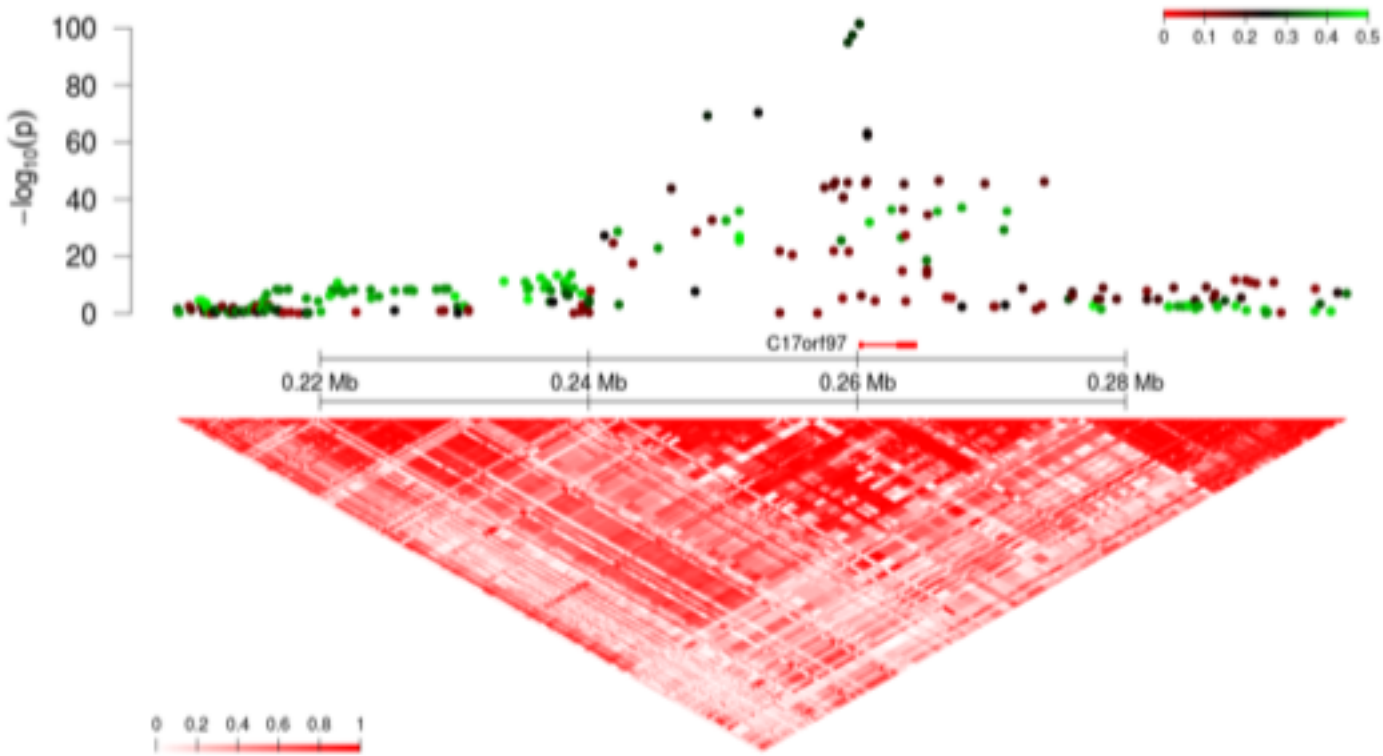
- **Resolution** - the region of the genome indicated by significant tests for a set of correlated markers in a GWAS
- Recall that we often consider a set of contiguous significant markers (a "skyscraper" on a Manhattan plot) to indicate the location of a single causal polymorphism (although it need 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



Review: 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 11: 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 I 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 I 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 I error α set the Bonferroni Type I 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 I error of 0.05, we require a test to have a p-value less than 0.0005 to be considered significant

Correcting for multiple tests III

- A False Discovery Rated (FDR) based approach (there are many variants!) uses the expected number of false positives to set (=control) the type I error
- For *N* tests and a specified Type I error, the FDR is defined in terms or the number of cases where the null hypothesis is rejected R:

$$FDR = \frac{N * \alpha}{R}$$

- Intuitively, the FDR is the proportion of cases where we reject the null hypothesis that are false positives
- We can estimate the FDR for a GWAS, e.g. say for *N*=100,000 tests and a Type I error of 0.05, we reject the null hypothesis 10,000 times, the FDR = 0.5
- FDR methods for controlling for multiple tests (e.g. Benjamini-Hochberg) set the Type I error to control the FDR to a specific level, say FDR=0.01 (what is the intuition at this FDR level?)

Correcting for multiple tests IV

- Since the lower the Type 1 error the lower the power of our test, if we set the Type 1 error too low due to a very large N, we might not get any hits even when there are clear causal polymorphisms (is this desirable!?)
- In general, a Bonferroni correction sets a lower overall GWAS Type I error than FDR approaches (what are the trade-offs and why would we choose one over the other?)
- Both Bonferroni and FDR approaches make the implicit assumption that all tests are independent (which we know not to be the case in GWAS!)
- A strategy that can produce a more accurate Bonferroni or FDR cutoff is to use a permutation approach (which we do not have time to cover in this course)
- Regardless of the approach, some correction for multiple tests is necessary to guard against a case where there are no true positives in the experiment, i.e. this is why we do not automatically assume the highest "peak" is a true positive (unless it is significant after a multiple test correction)

Statistical / experimental issues that affect True Positives: power 1

- Recall that *power* is defined as the probability of correctly rejecting the null hypothesis when it is false (incorrect)
- Also recall that we cannot control power directly because it depends on the true parameter value(s) that we do not know!
- Also recall that we can indirectly control power by setting our Type 1 error, where there is a trade-off between Type 1 error and power (what is this trade-off!?)
- There are also a number of issues that affect power that are a function of the GWAS experiment

Statistical / experimental issues that affect True Positives: power II

- Power tends to increase with the increasing size of the true effect of the genotype on phenotype (how is this quantified in terms of linear regression parameters?)
- Power tends to increase with increasing sample size *n*
- Power tends to increase as the Minor Allele Frequency (MAF) increases (why is this?)
- Power tends to increase as the LD between a causal polymorphism and the genotype marker being tested increases (i.e. as the correlation between the causal and marker genotype increase)
- Power also depends on other factors including the type of statistical test applied, etc.
- Can any of these be controlled?

Experimental issues that produce false positives

- Type I errors can produce a false positives (= places we identify in the genome as containing a causal polymorphism / locus that do not)
- However, there are experimental reasons why we can correctly reject the null hypothesis (= we do not make a Type I error) but we still get a false positive:
  - Cases of disequilibrium when there is no linkage
  - Genotyping errors
  - **Unaccounted for covariates**
  - There are others...

Quantile-Quantile (QQ) plots 1

- We will now introduce an essential tool for detecting the most problematic covariates (and can be used to diagnose many other problems!): a Quantile-Quantile (QQ) plot
- While the definition of a QQ-plot is complex, you will see that how we generate a QQ-plot is easy!
- We will demonstrate the value of a QQ plot for detecting the often problematic variable: population structure
- In general, whenever you perform a GWAS, you should construct a QQ plot (!!) and always include a QQ plot in your publication

Quantile-Quantile (QQ) plots II

- Consider a random variable with a continuous probability distribution
- **quantile** - regular, equally spaced intervals of a random variable that divide the random variable into units of equal distribution
- A Quantile-Quantile (QQ) plot (in general) plots the observed quantiles of one distribution versus another OR plots the observed quantiles of a distribution versus the quantiles of the ideal distribution
- We will use a QQ plot to plot our the quantile distribution of observed p-values (on the y-axis) versus the quantile distribution of expected p-values (what distribution is this!?)

Quantile-Quantile (QQ) plots III

- How to construct a QQ plot for a GWAS:
  - If you performed *N* tests, take the -log (base 10) of each of the p-values and put them in rank order from smallest to largest
  - Create a vector of *N* values evenly spaces from 1 to 1 / *N* (how do we do this?), take the -log of each of these values and rank them from smallest to largest
  - Take the pair of the smallest of values of each of these lists and plot a point on an x-y plot with the observed -log p-value on the y-axis and the spaced -log value on the x-axis
  - Repeat for the next smallest pair, for the next, etc. until you have plotted all *N* pairs in order

That's it for today

- Next lecture (Tues, March 28), we will covariates in GWAS and QQ plots!