Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

Lecture 18: Introduction to Covariates (and QQ plots)

Jason Mezey

March 28, 2023 (T) 8:05-9:20

Announcements

- Schedule Reminders:
  - We will NOT have lecture Thurs (March 30)
  - We will NOT have computer labs this week Thurs / Fri (March 30-31)
  - We will NOT have lectures OR computer labs (or office hours) next week (April 3-7) = Cornell, Ithaca Spring break
- Midterm Exam starts Weds, March 29 (!!) available in the morning on CMS and must be uploaded to CMS by 11:59pm Fri, March 31 (!!)

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!!).

Exam peek (genotypes)

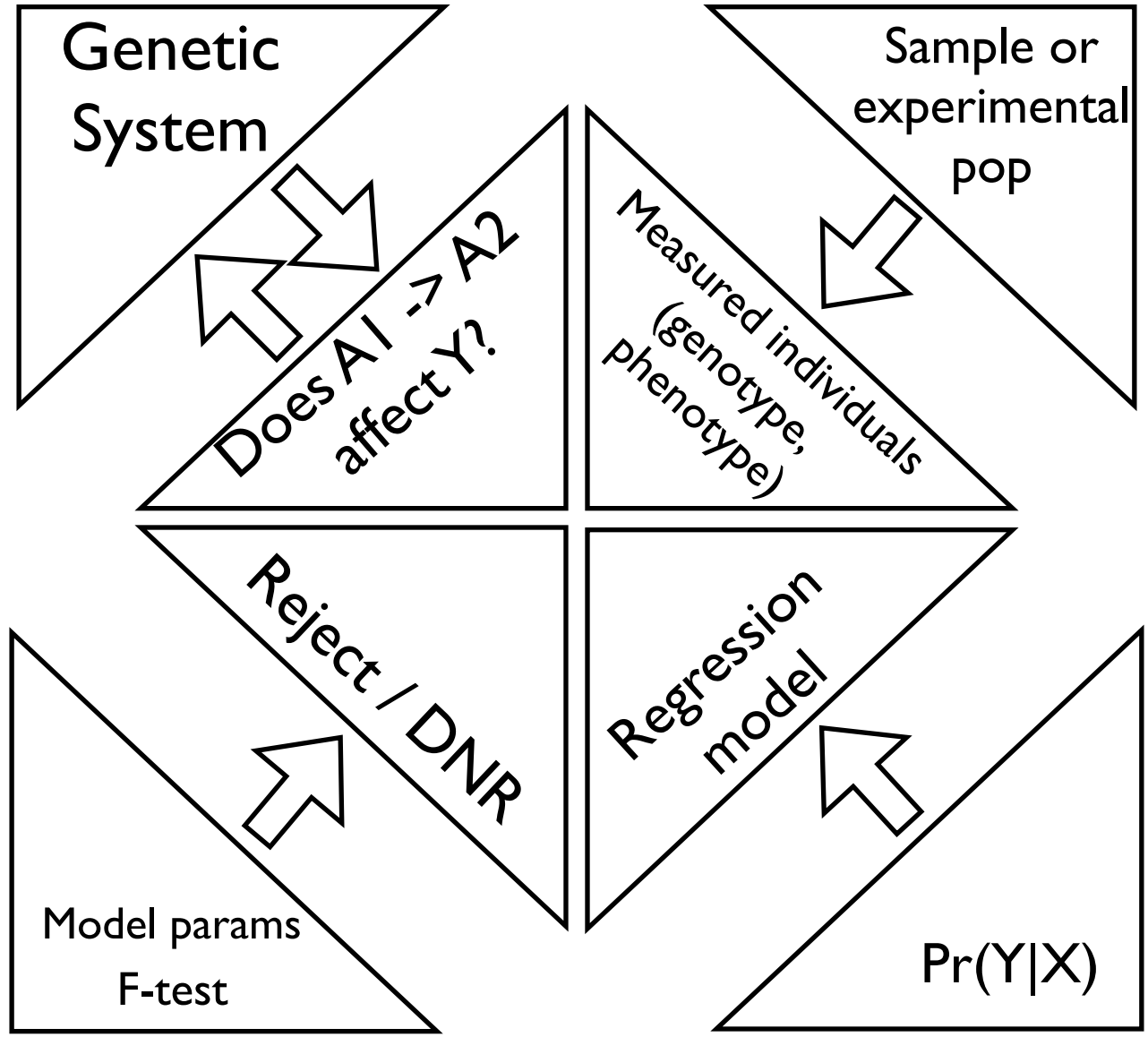
```
t,c,c,g,c,t,t,t,a,a,g,a,t,g,a,g,a,g,a,a,g,c,c,a,c,
c,t,c,g,g,c,t,c,t,a,c,a,t,g,c,c,c,a,g,c,t,g,c,c,c,
t,t,t,t,c,c,g,g,g,t,c,g,g,g,t,g,a,c,t,g,t,a,t,c,c,
g,c,c,g,c,a,c,g,t,c,t,g,c,g,c,a,t,a,g,c,c,t,g,t,a,
g,g,t,t,t,a,g,t,g,a,a,g,c,t,g,a,c,t,g,a,t,t,a,a,t,
c,g,g,a,t,g,t,t,a,t,c,t,t,g,a,c,g,t,g,a,g,a,t,t,c,
g,g,a,g,a,c,c,c,c,a,g,g,g,c,t,t,c,c,c,a,g,g,g,t,c,
g,a,a,a,t,c,c,t,g,t,t,c,g,t,a,t,c,t,c,t,g,t,a,t,t,
a,a,t,g,c,c,g,a,g,t,a,t,a,g,a,a,g,c,c,t,t,c,g,g,c,
t,c,t,a,t,g,a,g,t,a,t,t,t,c,c,t,g,t,a,t,c,c,a,g,c,
t,t,g,c,g,g,c,t,c,c,g,t,a,t,t,a,t,c,a,c,c,t,t,c,t,
a,c,t,a,t,c,g,g,t,t,g,t,c,a,g,t,g,a,a,a,a,a,g,g,c,
c,c,g,c,g,a,c,a,c,g,a,a,g,a,t,c,g,t,c,g,a,t,a,g,g,
t,a,t,t,a,t,g,c,c,t,c,c,a,c,t,g,a,a,t,t,g,t,g,c,t,
c,c,c,t,a,g,g,c,a,g,c,c,g,g,g,a,t,g,c,t,c,t,t,c,g,
```

And the genotype file will have a “header” (1st row with names of SNPs!)

Summary of lecture 18: Introduction to Covariates

- Last lecture, we discussed statistical and experimental issues impacting the success of a GWAS!
- Today we will continue this discussion by discussing covariates and QQ plots!

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:

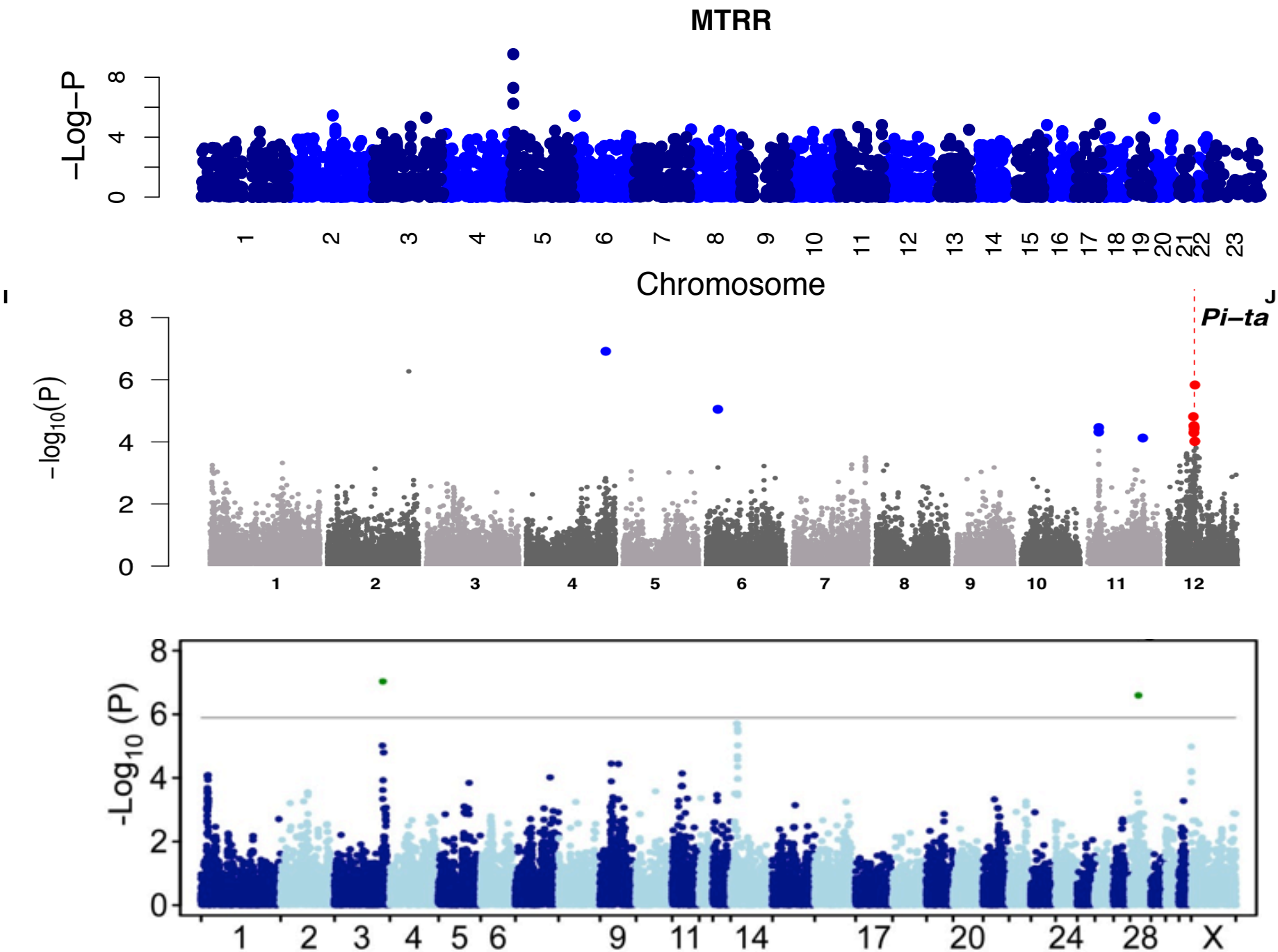
$$A_1 \rightarrow A_2 \Rightarrow \Delta Y | Z$$

- 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: Interpreting "hits" from a GWAS analysis

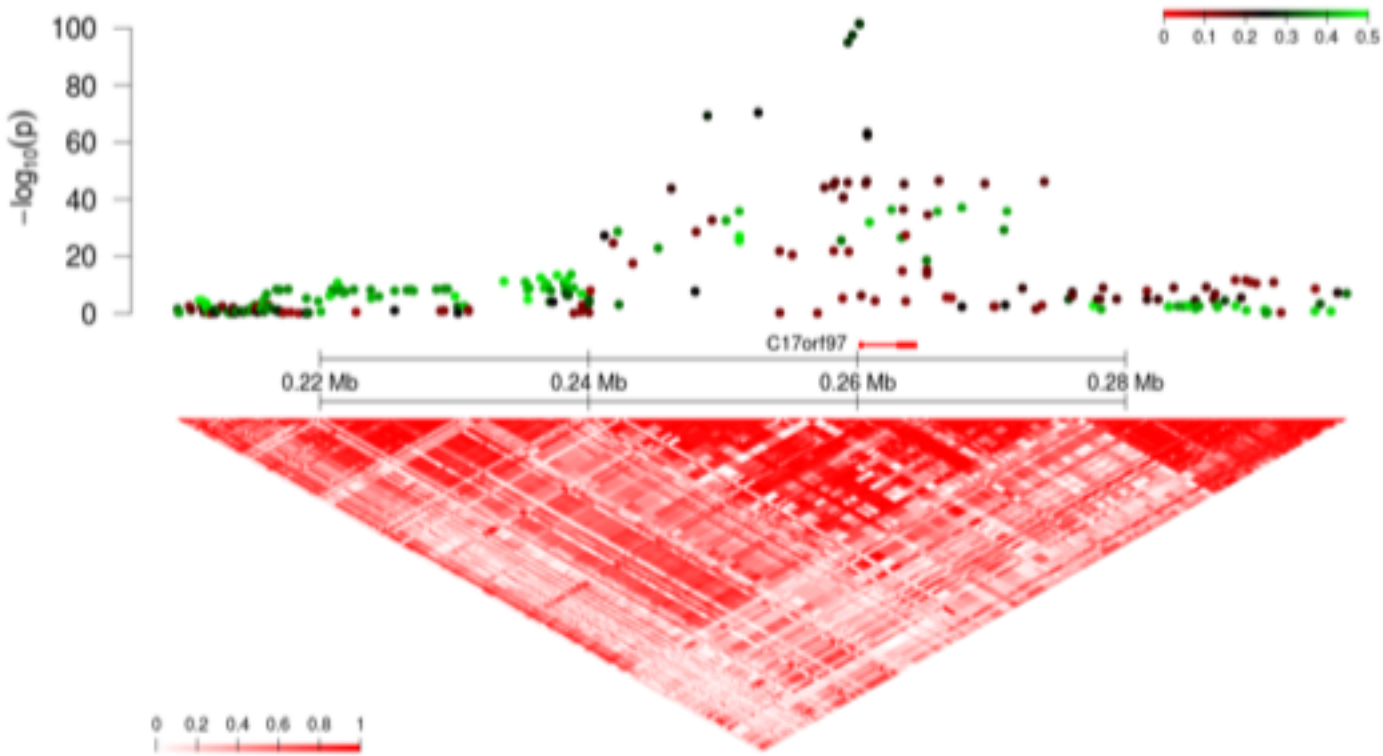
- **Resolution** - the region of the genome indicated by significant tests for a set of correlated markers in a GWAS
- Recall that we often consider a set of contiguous significant markers (a "skyscraper" on a Manhattan plot) to indicate the location of a single causal polymorphism (although it need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



Review: Issues for 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)

Review: 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*

Review: 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

Review: Correcting for multiple tests

- 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

Review: Correcting for multiple tests

- 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?)

Review: 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

Review: 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...

Introduction to covariates I

- Recall that in a GWAS, we are considering the following regression model and hypotheses to assess a possible association for every marker with the phenotype

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon$$

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

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

- Also recall that with these hypotheses we are actually testing:

$$H_0 : Cov(Y, X_a) = 0 \cap Cov(Y, X_d) = 0$$

$$H_A : Cov(Y, X_a) \neq 0 \cup Cov(Y, X_d) \neq 0$$

Introduction to covariates II

- Let's consider these two cases:

- For the first, the marker is not correlated with a causal polymorphism but the factor is correlated with BOTH the phenotype and the marker such that a test of the marker using our framework **will produce a false positive** (!!):

$$Cov(Y, X_z) \neq 0$$

$$Cov(X_a, X_z) \neq 0$$

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

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

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon$$

- For the second, the marker is correlated with a causal polymorphism and while the factor is correlated with the phenotype but not the marker, a test of the marker in our framework will model the effect of the factor in our error term (**which will reduce power**!):

$$Cov(Y, X_z) \neq 0$$

$$Cov(X_a, X_z) = 0$$

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon_{X_z}$$

$$\epsilon_{X_z} = X_z\beta_z + \epsilon$$

$$\epsilon \sim N(0, \sigma^2_\epsilon)$$

Modeling covariates I

- Therefore, if we have a factor that is correlated with our phenotype and we do not handle it in some manner in our analysis, we risk producing false positives AND/OR reduce the power of our tests!
- The good news is that, assuming we have measured the factor (i.e. it is part of our GWAS dataset) then we can incorporate the factor in our model as a *covariate(s)*:

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + X_{z,1}\beta_{z,1} + X_{z,2}\beta_{z,2} + \epsilon$$

- The effect of this is that we will estimate the covariate model parameter and this will account for the correlation of the factor with phenotype (such that we can test for our marker correlation without false positives / lower power!)

Modeling covariates II

- How do we perform inference with a covariate in our regression model?
- We perform MLE the same way (!!) our X matrix now simply includes extra columns, one for each of the additional covariates, where for the linear regression we have:

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

- We perform hypothesis testing the same way (!!) with a slight difference: our LRT includes the covariate in both the null hypothesis and the alternative (and therefore two different X matrices!), but we are testing the same null hypothesis:

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

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

Modeling covariates IV

- First, determine the predicted value of the phenotype of each individual under the null hypothesis (how do we set up **x**?):

$$\hat{y}_{i,\hat{\theta}_0} = \hat{\beta}_{\mu,\hat{\theta}_0} + \sum_{j=1} x_{i,z,j}\hat{\beta}_{z,\hat{\theta}_0,j}$$

- Second, determine the predicted value of the phenotype of each individual under the alternative hypothesis (set up **x**?):

$$\hat{y}_{i,\hat{\theta}_1} = \hat{\beta}_{\mu,\hat{\theta}_1} + x_{i,a}\hat{\beta}_{a,\hat{\theta}_1} + x_{i,d}\hat{\beta}_{d,\hat{\theta}_1} + \sum_{j=1} x_{i,z,j}\hat{\beta}_{z,\hat{\theta}_1,j}$$

- Third, calculate the "Error Sum of Squares" for each:

$$SSE(\hat{\theta}_0) = \sum_{i=1}^{n}(y_i - \hat{y}_{i,\hat{\theta}_0})^2 \qquad SSE(\hat{\theta}_1) = \sum_{i=1}^{n}(y_i - \hat{y}_{i,\hat{\theta}_1})^2$$

- Finally, we calculate the F-statistic with degrees of freedom [2, n-3] (why two and n-#params degrees of freedom?):

$$F_{[2,n-\#(\hat{\theta}_1)]}(\mathbf{y}, \mathbf{x_a}, \mathbf{x_d}) = \frac{\frac{SSE(\hat{\theta}_0)-SSE(\hat{\theta}_1)}{2}}{\frac{SSE(\hat{\theta}_1)}{n-\#(\hat{\theta}_1)}}$$

Modeling covariates V

- Thus, for testing the null hypothesis in a linear regression, we can construct an F-test using a slightly different formula:

$$SSE(\hat{\theta}_0) = \sum_{i=1}^{n} (y_i - \hat{y}_{i,\hat{\theta}_0})^2$$

$$SSE(\hat{\theta}_1) = \sum_{i=1}^{n} (y_i - \hat{y}_{i,\hat{\theta}_1})^2$$

$$F_{[2,n-\#(\hat{\theta}_1)]}(\mathbf{y}, \mathbf{x_a}, \mathbf{x_d}) = \frac{\frac{SSE(\hat{\theta}_0) - SSE(\hat{\theta}_1)}{2}}{\frac{SSE(\hat{\theta}_1)}{n-\#(\hat{\theta}_1)}}$$

- For the null hypotheses we are testing, once you calculate this F-statistic, you compare to an F-distribution with 2 and n - #(alternative hypothesis parameters) degrees of freedom
- The "2" df in the numerator comes from the #(alternative hypothesis model parameters) - #(null hypothesis model parameters)
- Note that our previous formula for an F-statistic can be represented this way as well (!!)

Modeling covariates VI

- Say you have GWAS data (a phenotype and genotypes) and your GWAS data also includes information on a number of covariates, e.g. male / female, several different ancestral groups (different populations!!), other risk factors, etc.
- First, you need to figure out how to code the XZ in each case for each of these, which may be simple (male / female) but more complex with others (where how to code them involves fuzzy rules, i.e. it depends on your context!!)
- Second, you will need to figure out which to include in your analysis (again, fuzzy rules!) but a good rule is if the parameter estimate associated with the covariate is large (=significant individual p-value) you should include it!
- There are many ways to figure out how to include covariates (again a topic in itself!!) - next lecture we will provide an (important!) example: population structure

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

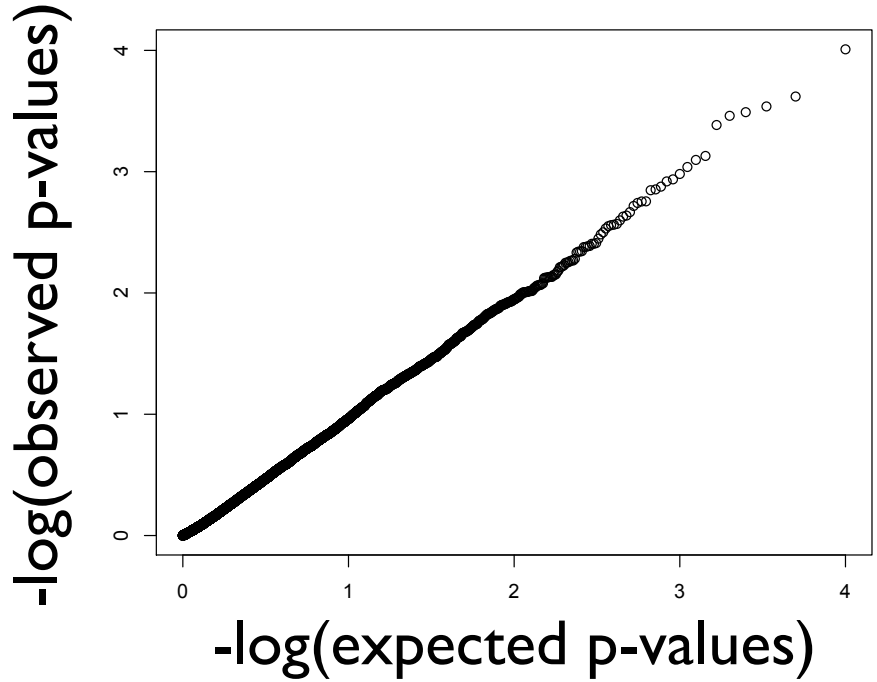
- Consider a random variable with a continuous probability distribution
- **quantile** - regular, equally spaced intervals of a random variable that divide the random variable into units of equal distribution
- A Quantile-Quantile (QQ) plot (in general) plots the observed quantiles of one distribution 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

Quantile-Quantile (QQ) plots III

- In an ideal GWAS case where there ARE NO causal polymorphisms, your QQ plot will be a line:

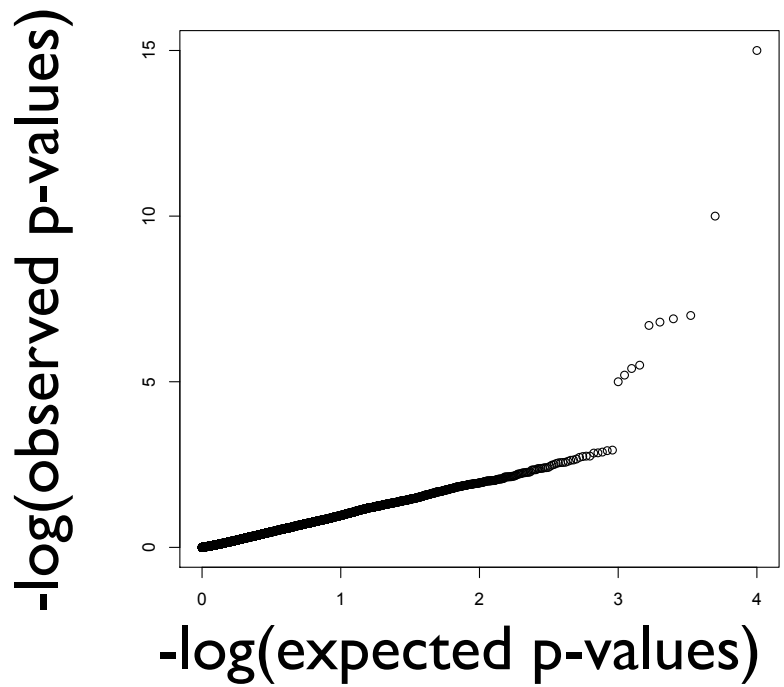


- The reason is that we will observe a uniform distribution of p-values from such a case and in our QQ we are plotting this observed distribution of p-value versus the expected distribution of p-values: a uniform distribution (where both have been -log transformed)

- Note that if you GWAS analysis is correct but you do not have enough power to detect positions of causal polymorphisms, this will also be your result (!!), i.e. it is a way to assess whether you can detect any hits in your GWAS (!!)

Quantile-Quantile (QQ) plots IV

- In an ideal GWAS case where there ARE causal polymorphisms, your QQ plot will be a line with a tail (!!):

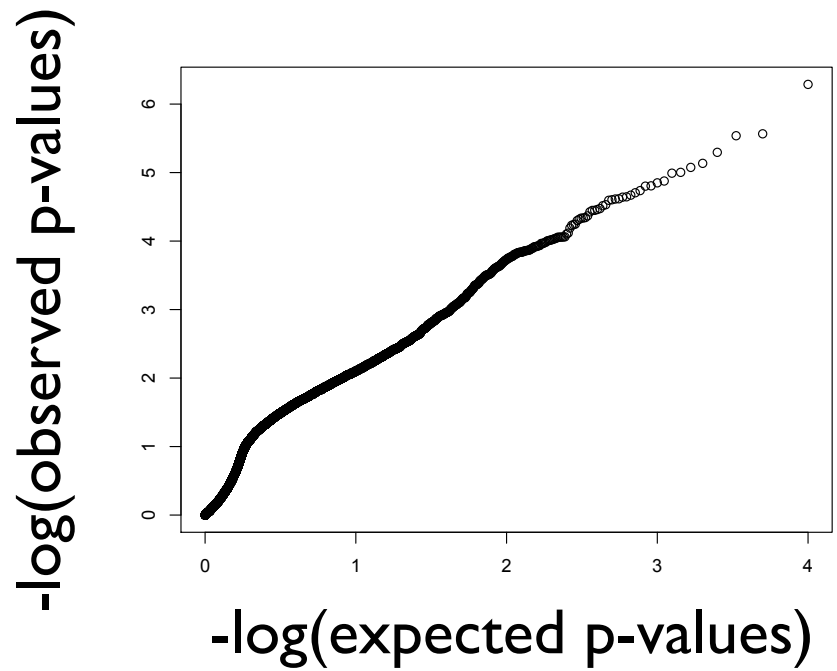


- This happens because most of the p-values observed follow a uniform distribution (i.e. they are not in LD with a causal polymorphism so the null hypothesis is correct!) but the few that are in LD with a causal polymorphism will produce significant p-values (extremely low = extremely high -log(p-values)) and these are in the "tail"

- This is ideally how you want your QQ-plot to look - if it does, you are in good shape!

Quantile-Quantile (QQ) plots V

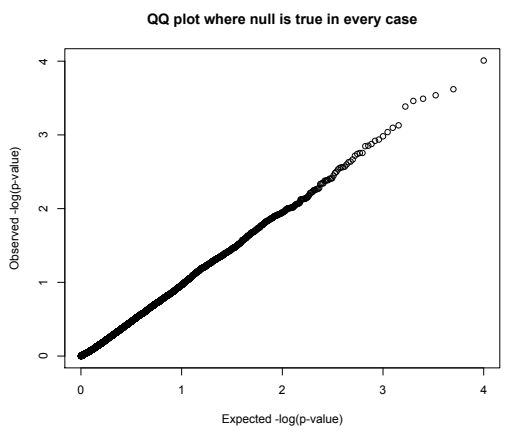
- In practice, you can find your QQ plot looks different than either the "null GWAS" case or the "ideal GWAS" case, for example:

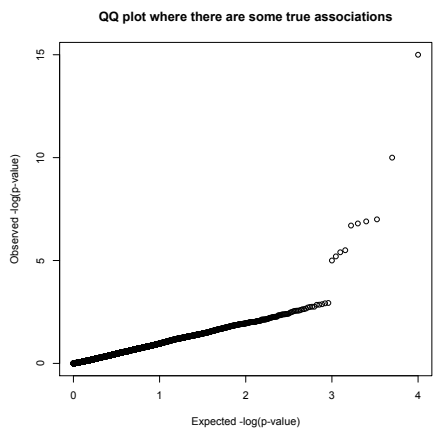


- This indicates that something is wrong (!!!!) and if this is the case, you should not interpret any of your significant p-values as indicating locations of causal polymorphisms (!!!!)
- Note that this means that you need to find an analysis strategy such that the result of your GWAS produces a QQ plot that does NOT look like this (note that this takes experience and many tools to do consistently!)
- Also note that unaccounted for covariates can cause this issue and the most frequent culprit is unaccounted for population structure

Important (!!): when to use / how to interpret QQ diagnostics

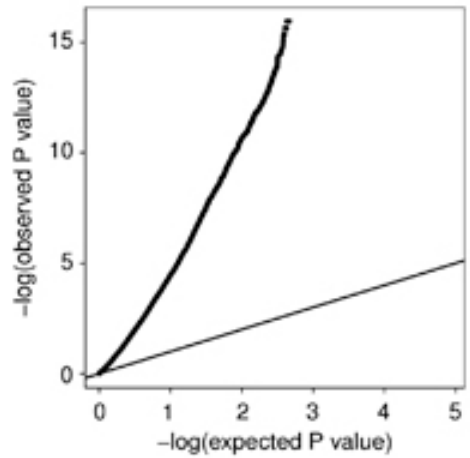
- In a GWAS (i.e., when you have a single phenotype and you are considering the impact of MANY genotypes!) always use a QQ and interpret two cases (i.e., all on 45 deg line or most on 45 deg line with "tail" as an indicator to interpret analysis results (otherwise there is a problem!)





- Caveat: there can be exceptions… but make sure you understand when these occur and why (!!)

- In analyses with MANY phenotypes and a single genotype, it is very possible that the genotype impacts many phenotypes producing way more significant tests and a QQ that would NOT be acceptable for GWAS but is FINE for assessing a single genotype impact on many phenotypes:



- Plotting a QQ can still be useful in these cases (=recommended!)

That's it for today

- See you April 11 after Spring break!