Quantitative Genomics and Genetics

BioCB 4830/6830; PBSB.5201.03

Lecture 20: Population Structure in GWAS

Jason Mezey

April 9, 2024 (T) 8:40-9:55

Announcements

- Your midterm (!!) STARTS TODAY (by 11AM) = once the exam has started, DO NOT COMMUNICATE WITH ANYONE IN ANY WAY ABOUT ANY ASPECT OF THE EXAM (work on your own!) - the only exception is myself, Beulah, Sam (we will answer clarifying questions).
- No office hours tomorrow (!!)

Summary of lecture 20: Population Structure in GWAS

- Last lecture, we continued our discussion GWAS analysis issues, including multiple test correction, covariates and QQ plots!
- Today we will continue our discussion of using covariates (and QQ plots) by discussing a critical covariates in GWAS: population structure!

Review: True and False Positive Trade-offs 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: Experimental issues that produce false positives

- Type 1 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 1 error) but we still get a false positive:
  - Cases of disequilibrium when there is no linkage
  - Genotyping errors
  - **Unaccounted for covariates**
  - There are others...

Review: 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:

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

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

Review: 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**!):

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

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

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

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

$$\epsilon \sim N(0, \sigma_\epsilon^2)$$

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!!)

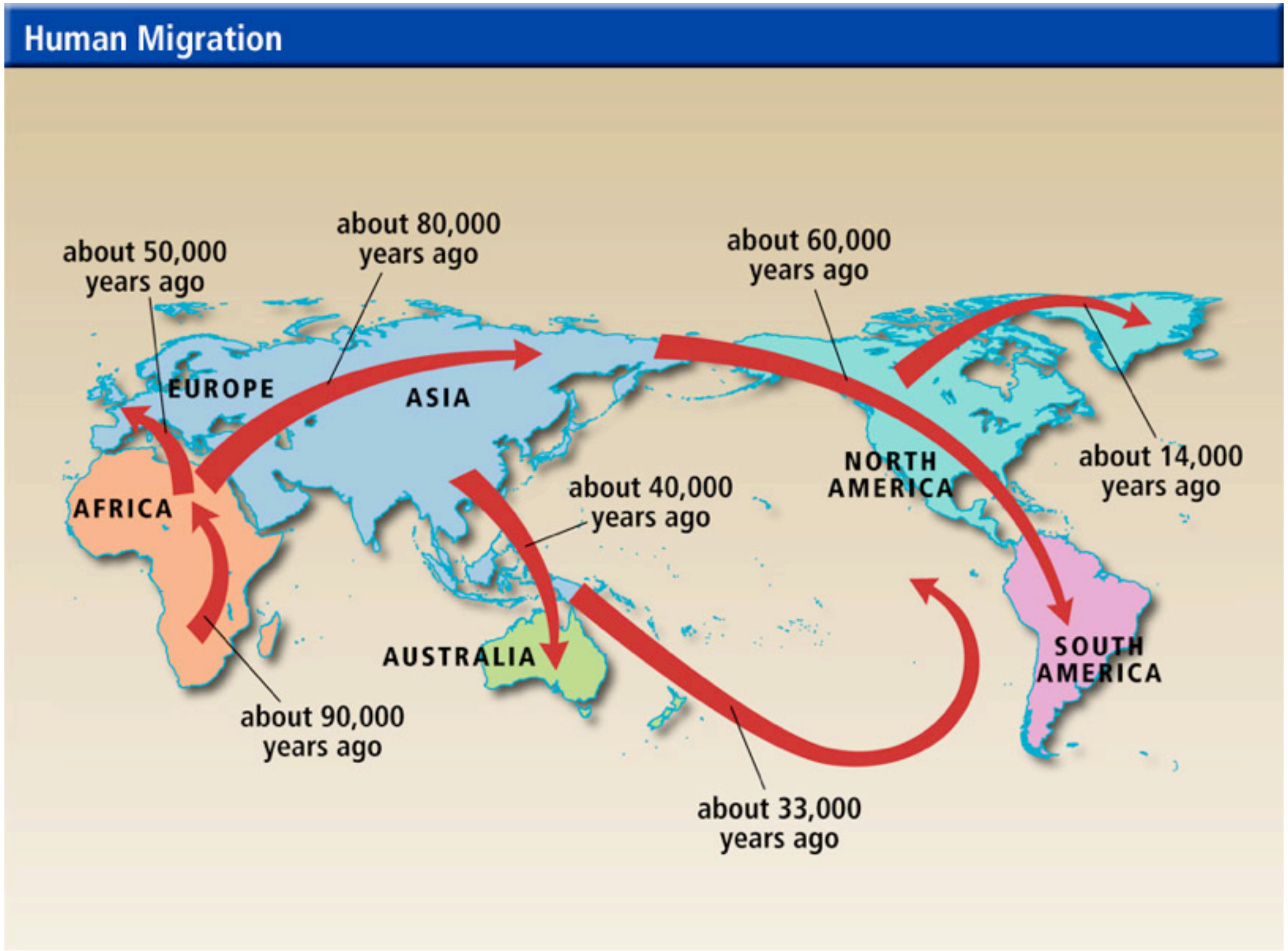
- Second, you will need to figure out which to include in your analysis (again, fuzzy rules!) but a good rule is if the parameter estimate associated with the covariate is large (=significant individual p-value) you should include it!

- 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

Covariate modeling example: population structure

- "Population structure" or "stratification" is a case where a sample includes groups of people that fit into two or more different ancestry groups (fuzzy def!)
- Population structure is often a major issue in GWAS where it can cause lots of false positives if it is not accounted for in your model
- Intuitively, you can model population structure as a covariate if you know:
  - How many populations are represented in your sample
  - Which individual in your sample belongs to which population
- QQ plots are good for determining whether there may be population structure
- "Clustering" techniques are good for detecting population structure and determining which individual is in which population (=ancestry group)

Origin of population structure



© Sarver World Cultures

People geographically separate through migration and then the set of alleles present in the population evolves (=changes) over time

Why might (unaccounted for) structure be a problem in a GWAS?

- Even if you had a case where there were NO causal polymorphisms for a phenotype, you can get false positives if:
  - If you have more than one population in your sample (that you do not model with a covariate)
  - If these populations differ in frequencies of genotypes at a subset of measured genotypes / polymorphisms
  - If these populations differ in the mean value of the phenotype
- In such a case, every genotype where an MAF is different between the populations would be expected to produce a low p-value (=biological false positives!)
- Note: if you can “learn” (or know) the population information for your data, you can model this as a covariate and you (may) be able to correct this problem

Modeling population structure as a covariate (intuition)

- If you can determine which individual is in which pop and define random variables for pop assignment, e.g. for two populations include single covariate by setting, $X_{Z,1}$(pop1) = 1, $X_{Z,1}$(pop2) = 0 (generally less optimal but can be used!)
- Use one of these approaches to model a covariate in your analysis, i.e. for every genotype marker that you test in your GWAS:

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

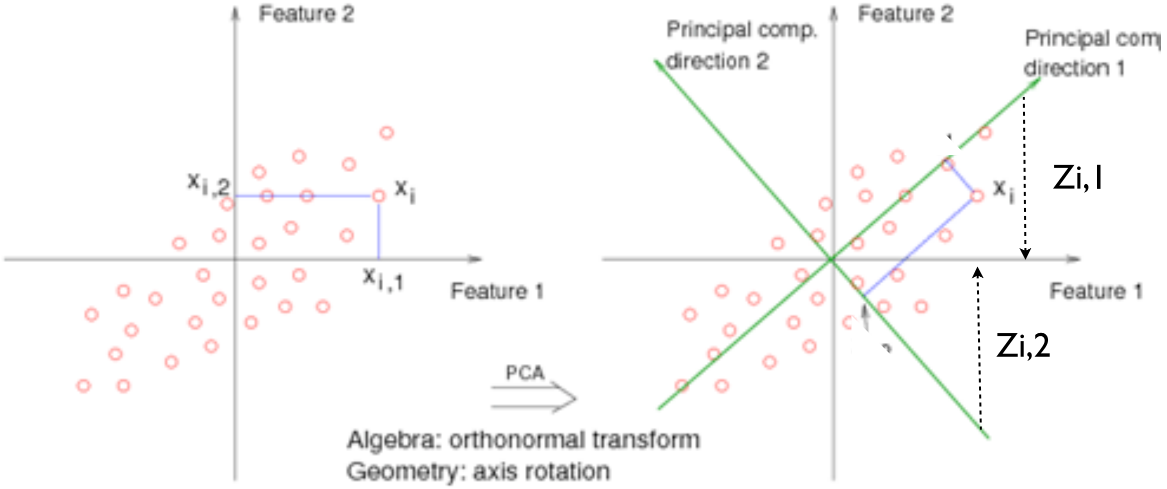
- How do we tell if our covariate correction "worked" well enough that we should interpret the results of our analysis?

Learning unmeasured population factors

- To learn a population factor, analyze the genotype data

$$Data = \left[\begin{matrix} z_{11} & \dots & z_{1k} & y_{11} & \dots & y_{1m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{n1} & \dots & z_{nk} & y_{n1} & \dots & y_{nm} \end{matrix} \boxed{\begin{matrix} x_{11} & \dots & x_{1N} \\ \vdots & \vdots & \vdots \\ x_{11} & \dots & x_{nN} \end{matrix}}\right]$$

- Apply a Principal Component Analysis (PCA) where the "axes" (features) in this case are individuals and each point is a (scaled) genotype



- What we are interested in the projections (loadings) of the individual PCs on the axes (dotted arrows) on each of the individual axes, where for each, this will produce *n* (i.e. one value for each sample) value of a new independent (covariate) variable X*z*

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

Using the results of a PCA population structure analysis

- Once you have detected the populations (e.g. by eye in a PCA = fuzzy!) in your GWAS sample, set your independent variables equal to the loadings for each individual, e.g., for two pop covariates, set $X_{Z,1} = Z_1$, $X_{Z,2} = Z_2$

- You could also determine which individual is in which pop and define random variables for pop assignment, e.g. for two populations include single covariate by setting, $X_{Z,1}(pop1) = 1$, $X_{Z,1}(pop2) = 0$ (generally less optimal but can be used!)

- Use one of these approaches to model a covariate in your analysis, i.e. for every genotype marker that you test in your GWAS:

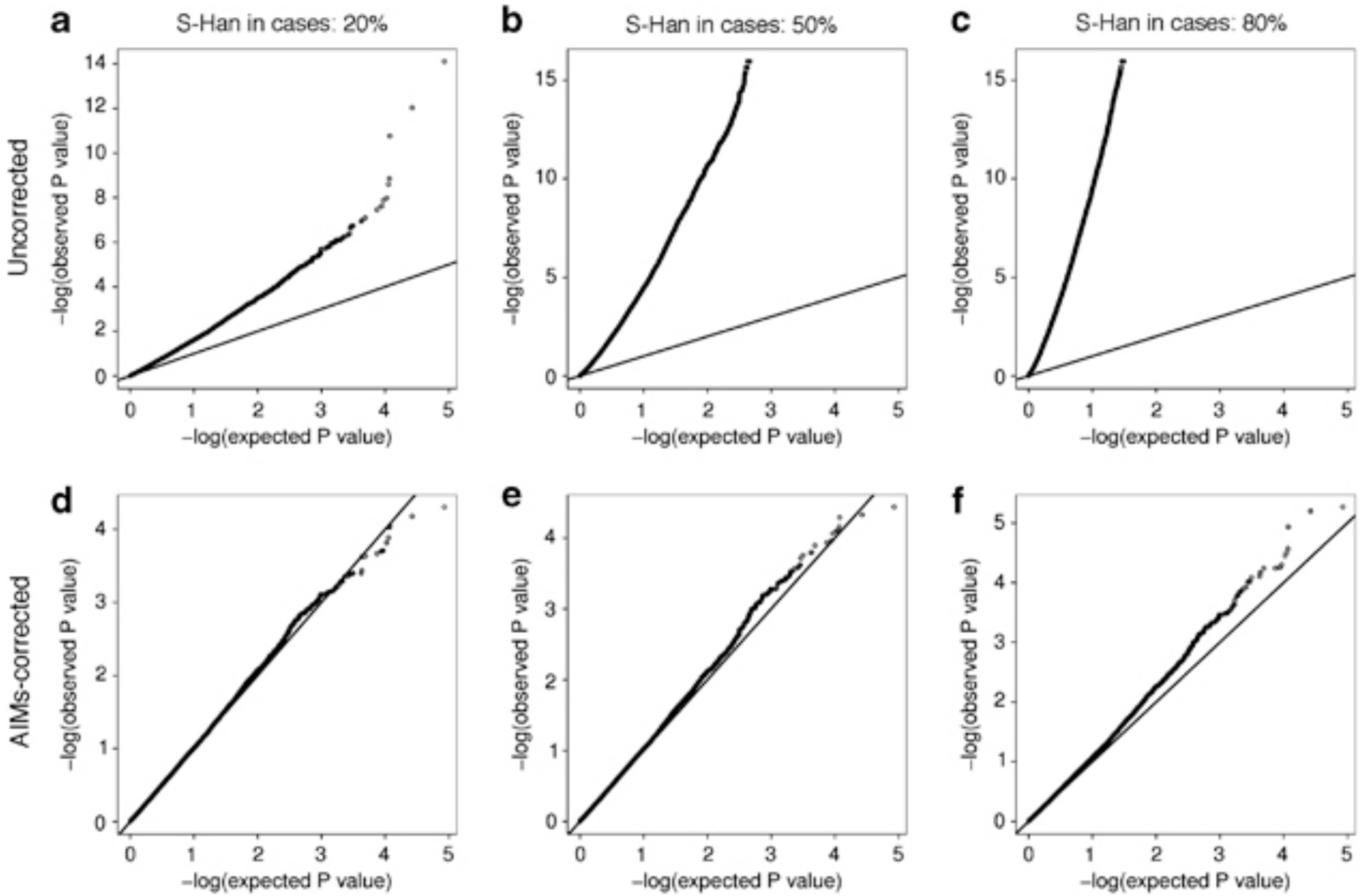
$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + X_{z,1}\beta_{z,1} + X_{z,2}\beta_{z,2} + \epsilon$$

- The goal is to produce a good QQ plot (what if it does not?)

Reminder: Quantile-Quantile (QQ) plots

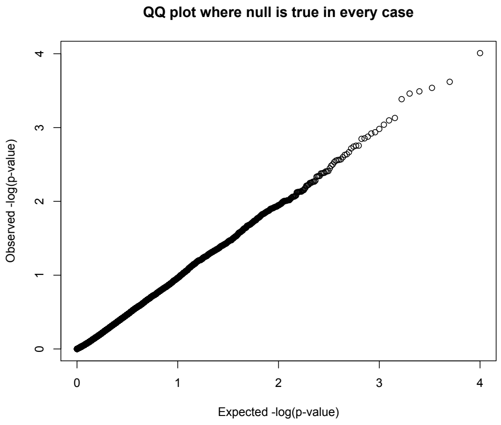
- We will now introduce an essential tool for detecting the most problematic covariates (and can be used to diagnose many other problems!): a Quantile-Quantile (QQ) plot
- While the definition of a QQ-plot is complex, you will see that how we generate a QQ-plot is easy!
- We will demonstrate the value 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

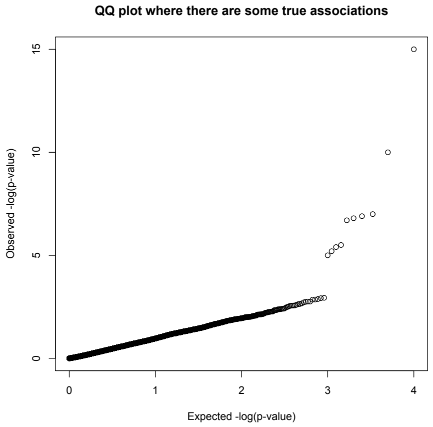
Before (top) and after including a population covariate (bottom)



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

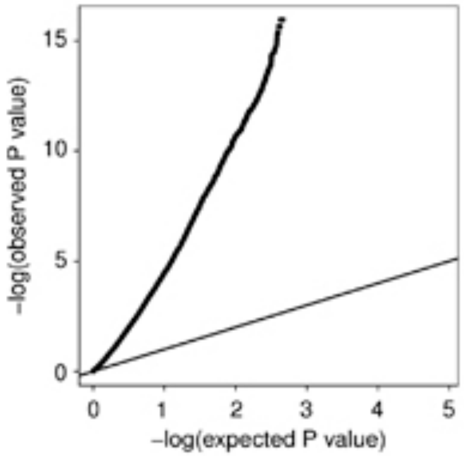
- In a GWAS (i.e., when you have a single phenotype and you are considering the impact of MANY genotypes!) always use a QQ and interpret two cases (i.e., all on 45 deg line or most on 45 deg line with "tail" as an indicator to interpret analysis results (otherwise there is a problem!)


QQ plot where null is true in every case


QQ plot where there are some true associations

- 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

- Next lecture, we will discuss minimal GWAS analysis steps (and begin our discussion of logistic regression) (!!)