Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

Lecture 19: Introduction to population structure

Jason Mezey
April 11, 2023 (T) 8:05-9:20

Announcements

- Homework #3 and #4 have been graded - grades will be released later today
- Midterm will be graded soon and key will be available soon
- You only have your Project and the Final Exam (and 4 computer labs) to go!
- Project (!!) will be available today (see next slides for more information)
- Looking ahead: you final:
    - Same basic format as the midterm
    - You will have to do a GWAS analysis by doing a linear regression with and without covariates AND a logistic regression with and without covariates (!!)

BTRY 4830/6830 & PBSB.5201.01
Quantitative Genomics and Genetics Spring 2023

Project - posted April 11

Due 11:59PM May 9

1 Introduction and instructions

The goal of the class project is for you to demonstrate what you have learned by performing a GWAS analysis on real data. To accomplish this, assume that you have been provided data by a collaborator who wants to identify positions of causal polymorphisms (loci). You will perform an in-depth analysis and write a report for your collaborator that explains your methods and results.

Instructions: While we provide some general guidelines for how to proceed below, the techniques you use to analyze the data and how you construct your report will be up to you. Do however note the following instructions (PLEASE READ THESE CAREFULLY!!):

(1) Your project must be uploaded by 11:59PM, May 9 - if it is late for any reason, standard grading policies apply.

(2) You are allowed to work together with other students in the class to analyze these data. However, note that turning in a report that describes exactly the same analyses as a fellow student is not a good strategy for getting a good grade. Also note that you must write your own report.

(3) This is an 'open book' assignment, such that you are allowed to use any resources online, in books, etc. You may also ask third-party (i.e. people not in the class) for suggestions on what analyses to perform but you cannot have a third-party do any of the analyses (or write any code for you!).

(4) You are also allowed to use any software or programming language that you would like as part of your analysis. However, we expect that some of the tasks will be performed in R (also note that you are welcome to use any packages, functions, etc. in R).

(5) Your final project will include at most three files a SINGLE report file (ideally a .pdf), a SINGLE file including all of your R code (ideally an .rmd file!) and / or commands or scripts you used to run other software packages, and IF YOU WANT a SINGLE, a pdf or html conversion of your .rmd. That is, for your R code, the best way to maximize your grade is to have well commented code that we can run from the command line. If you use other software for some of the tasks, a reasonable approach is to include commented out descriptions in your

code that provides details on how you ran the software, e.g. what parameters did you use, etc.

(6) The report file must be no more than 8 pages (single-sided), with NO MORE than 5 pages of text and NO MORE than 3 pages of figures / tables.

(7) For your report, you must describe what you did in detail (a good guide is have you provided enough detail such that someone reading your report could replicate what you have done?). You also need to describe the results you have obtained from your analysis. You may also wish to include some text to describe interpretations and conclusions that may be of interest to your collaborator, including statistical and possibly, biological interpretations. For your Figures and Tables, note that clarity and clear labels is a strategy for maximizing your grade.

(8) We will grade on two broad criteria: 1. the overall quality of the analyses / report, 2. the amount of effort put into your project. Note that 'effort' does not mean run many analyses without thinking carefully about why you are running them or how they fit together to provide a clear picture of results. A guide maximizing your grade on effort is to think carefully about how to produce the best possible report that you can and then put in as many hours as you wish to devote to the project accomplishing this objective (your effort level will be clear to us).

2 The experiment and data

The experiment: Among the recent large scale human genomics resources is Genetic European Variation in Health and Disease (gEUVADIS) - see the following links for relevant descriptions and information:

`http://www.internationalgenome.org/data-portal/data-collection/geuvadis/`

`https://www.nature.com/articles/nature12531`

with a samples from 4 different European populations (5 populations total). Each of these individuals were part of the 1000 Genomes project and their genomes were sequenced and analyzed to identify SNP geno- types. For expression profiling, lympoblastoid cell lines (LCL) were generated from each sample and mRNA levels were quantified through RNA sequencing.

Each of these gene expression measurements may be thought of as a phenotype and one can do a GWAS analysis on each individually, which is called an 'expression Quantitative Trait Locus' or 'eQTL' analysis, an unnecessarily fancy name for a GWAS when the phenotype is gene expression!

What you have been provided is a small subset of these data that are publicly available. Specifically, you have been provided 50,000 of the SNP genotypes for 344 samples from the CEU (Utah residents with European ancestry), FIN (Finns), GBR (British) and, TSI (Toscani) population. For these same individuals, you have also been provided the expression levels of five genes. You have also been provided information on the population and gender of each of these individuals, and information regarding the position of each gene and SNP in the genome.

The data: These have been provided to you in five total files: ‘phenotypes.csv’,‘genotypes.csv’, ‘covars.csv’, ‘gene_info.csv’,‘SNP_info.csv’.

‘phenotypes.csv’ contains the phenotype data for 344 samples and 5 genes.

‘genotypes.csv’ contains the SNP data for 344 samples and 50000 genotypes.

‘covars.csv’ contains the population origin and gender information for the 344 samples.

‘gene_info.csv’ contains information about each gene that was measured. The ‘chromosome’ column indicates the chromosome where the gene is located, ‘start’ marks the position in the chromosome where the region of the gene begins and ‘end’ marks the position where the region ends, ‘symbol’ contains the common gene name of the measured transcript and ‘probe’ contains the ids of the transcripts that match with the column names of the phenotype data.

‘SNP_info.csv’ contains the additional information on the genotypes and has four columns. The 1st column contains the chromosome number of each SNP, the 2nd column contains the physical position of the SNP on the chromosome, the 3rd column contains the abbreviation used to the ‘rsID’ = the name of each SNP in order.

3 Your assignment and hints for getting started

Your GWAS assignment is to find the position of as many causal polymorphisms as possible for the five expressed genes using the data (note that each ‘hit’ will potentially indicate an eQTL). You may / should use any and as many analysis approaches as you think that are useful to accomplish this goal. In your report, you will need to describe in detail what you did, why you did it, and describe results in a manner that your ‘non-statistical’ collaborator will be able to understand, e.g. explain your terms, provide interpretations, etc.

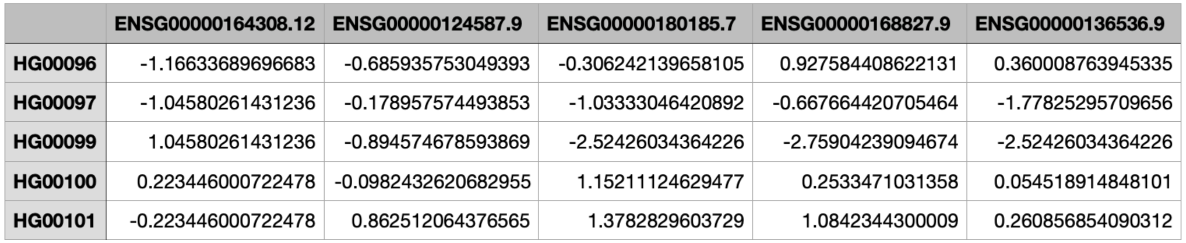
A few hints:

- Apply the applicable steps of a ‘minimum GWAS’ analysis.
- In your report, justify why you applied each individual step and statistical approach.
- In your report, provide a summary of your results and what they mean.
- You may want to consider going to various resources online (e.g. genecards, UCSC genome browser, dbSNP, many others) to incorporate biological information into your interpretation and hypotheses concerning what you may have found.
- Ask Mitch, Sam, and Jason for thoughts and ideas!

Good luck!

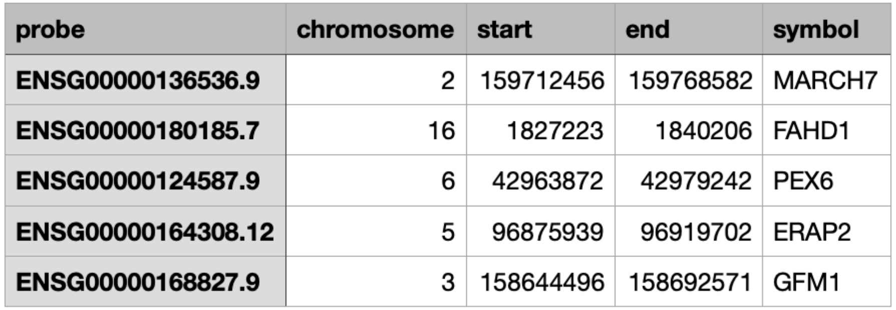
Phenotype data (gene expression = part of transcriptome)

phenotypes

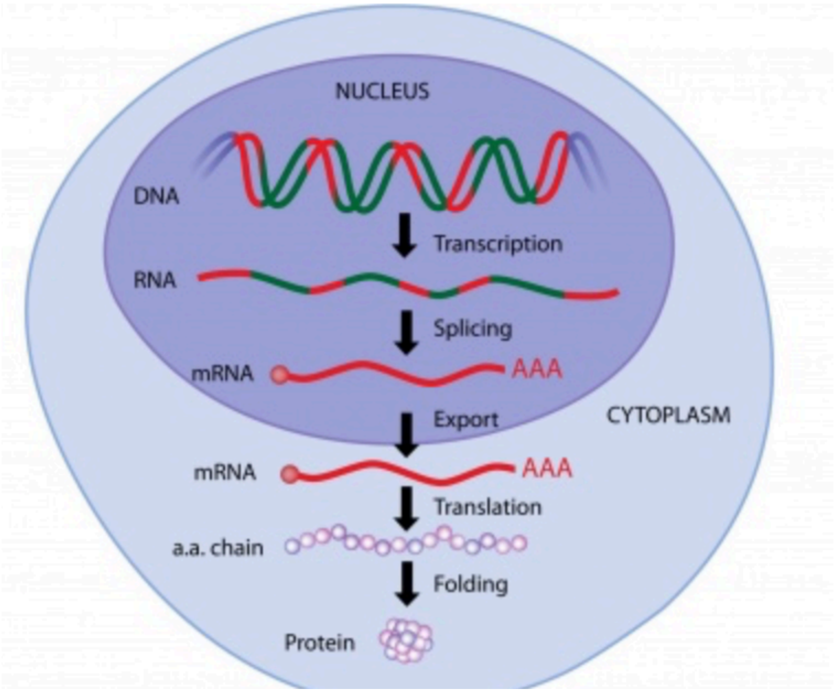


	ENSG00000164308.12	ENSG00000124587.9	ENSG00000180185.7	ENSG00000168827.9	ENSG00000136536.9
HG00096	-1.16633689696683	-0.685935753049393	-0.306242139658105	0.927584408622131	0.360008763945335
HG00097	-1.04580261431236	-0.178957574493853	-1.03333046420892	-0.667664420705464	-1.77825295709656
HG00099	1.04580261431236	-0.894574678593869	-2.52426034364226	-2.75904239094674	-2.52426034364226
HG00100	0.223446000722478	-0.0982432620682955	1.15211124629477	0.2533471031358	0.054518914848101
HG00101	-0.223446000722478	0.862512064376565	1.3782829603729	1.0842344300009	0.260856854090312

gene_info



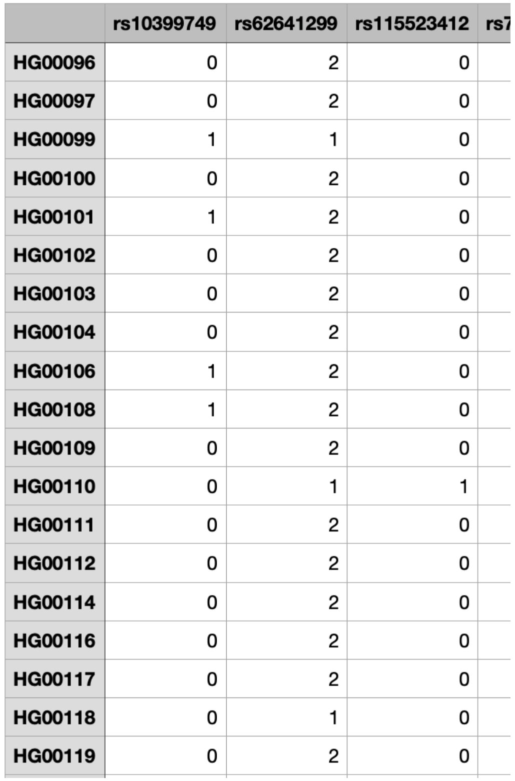
probe	chromosome	start	end	symbol
ENSG00000136536.9	2	159712456	159768582	MARCH7
ENSG00000180185.7	16	1827223	1840206	FAHD1
ENSG00000124587.9	6	42963872	42979242	PEX6
ENSG00000164308.12	5	96875939	96919702	ERAP2
ENSG00000168827.9	3	158644496	158692571	GFM1



intragen-genomics.com

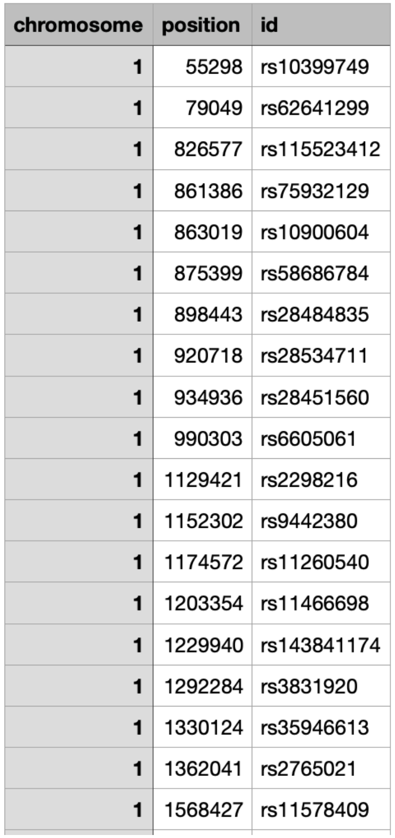
Phenotypes (5 total!) are expression level of 5 genes measured in (immortalized) LCL cells derived from individuals in the (original)1000 Genome project (so GWAS in this case is an eQTL study!)

Genotype data (SNPs across the genome)



	rs10399749	rs62641299	rs115523412	rs7
HG00096	0	2	0	
HG00097	0	2	0	
HG00099	1	1	0	
HG00100	0	2	0	
HG00101	1	2	0	
HG00102	0	2	0	
HG00103	0	2	0	
HG00104	0	2	0	
HG00106	1	2	0	
HG00108	1	2	0	
HG00109	0	2	0	
HG00110	0	1	1	
HG00111	0	2	0	
HG00112	0	2	0	
HG00114	0	2	0	
HG00116	0	2	0	
HG00117	0	2	0	
HG00118	0	1	0	
HG00119	0	2	0	

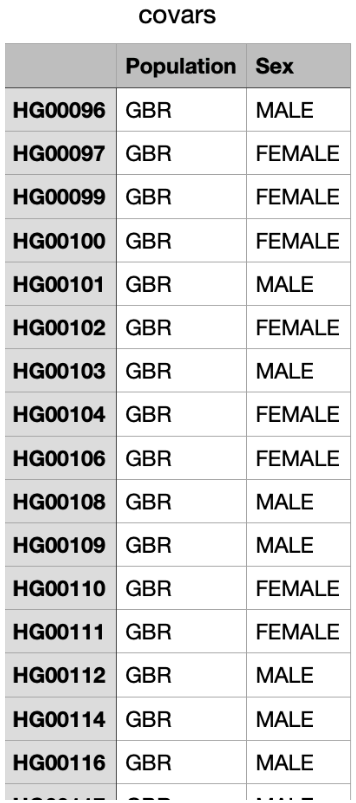
SNP_info



chromosome	position	id
1	55298	rs10399749
1	79049	rs62641299
1	826577	rs115523412
1	861386	rs75932129
1	863019	rs10900604
1	875399	rs58686784
1	898443	rs28484835
1	920718	rs28534711
1	934936	rs28451560
1	990303	rs6605061
1	1129421	rs2298216
1	1152302	rs9442380
1	1174572	rs11260540
1	1203354	rs11466698
1	1229940	rs143841174
1	1292284	rs3831920
1	1330124	rs35946613
1	1362041	rs2765021
1	1568427	rs11578409

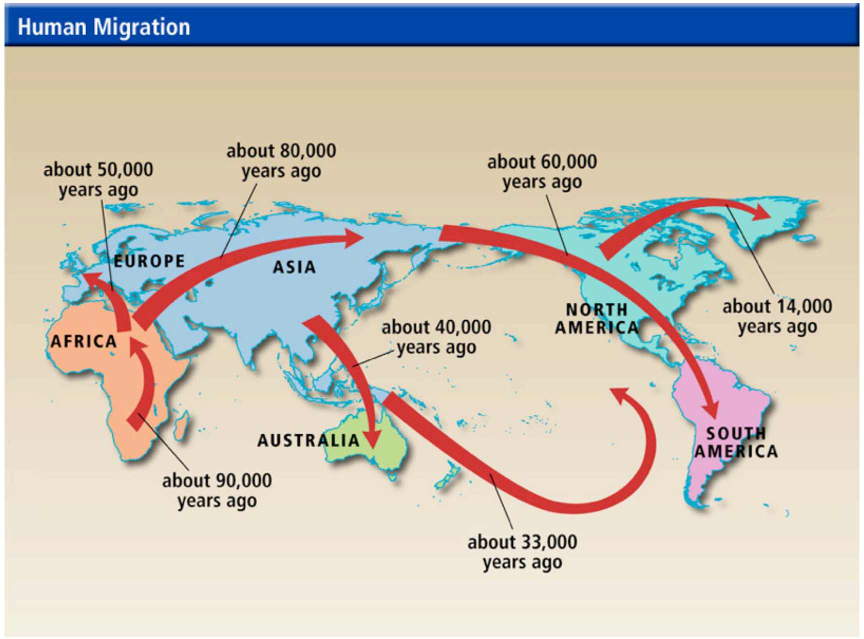
SNPs (original alleles are a, c, g, t) have been coded data as 0 (=homozygote), 1 (=heterozygote), 2 (= other homozygote) with "rsID" for identity (and tells you where the SNP is located in the ref genome)

Other (Covar) data (gender and population)



covars

	Population	Sex
HG00096	GBR	MALE
HG00097	GBR	FEMALE
HG00099	GBR	FEMALE
HG00100	GBR	FEMALE
HG00101	GBR	MALE
HG00102	GBR	FEMALE
HG00103	GBR	MALE
HG00104	GBR	FEMALE
HG00106	GBR	FEMALE
HG00108	GBR	MALE
HG00109	GBR	MALE
HG00110	GBR	FEMALE
HG00111	GBR	FEMALE
HG00112	GBR	MALE
HG00114	GBR	MALE
HG00116	GBR	MALE



Other data include information about the gender of each individual in the sample and where their ancestry group (GBR, FIN, CEU, TSI codes indicate different populations)

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**!):

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

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

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

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

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

Review: 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!)

Review: 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:

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

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

Review: 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?):

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

Review: 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 (!!)

Review: 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 X_z 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!!)

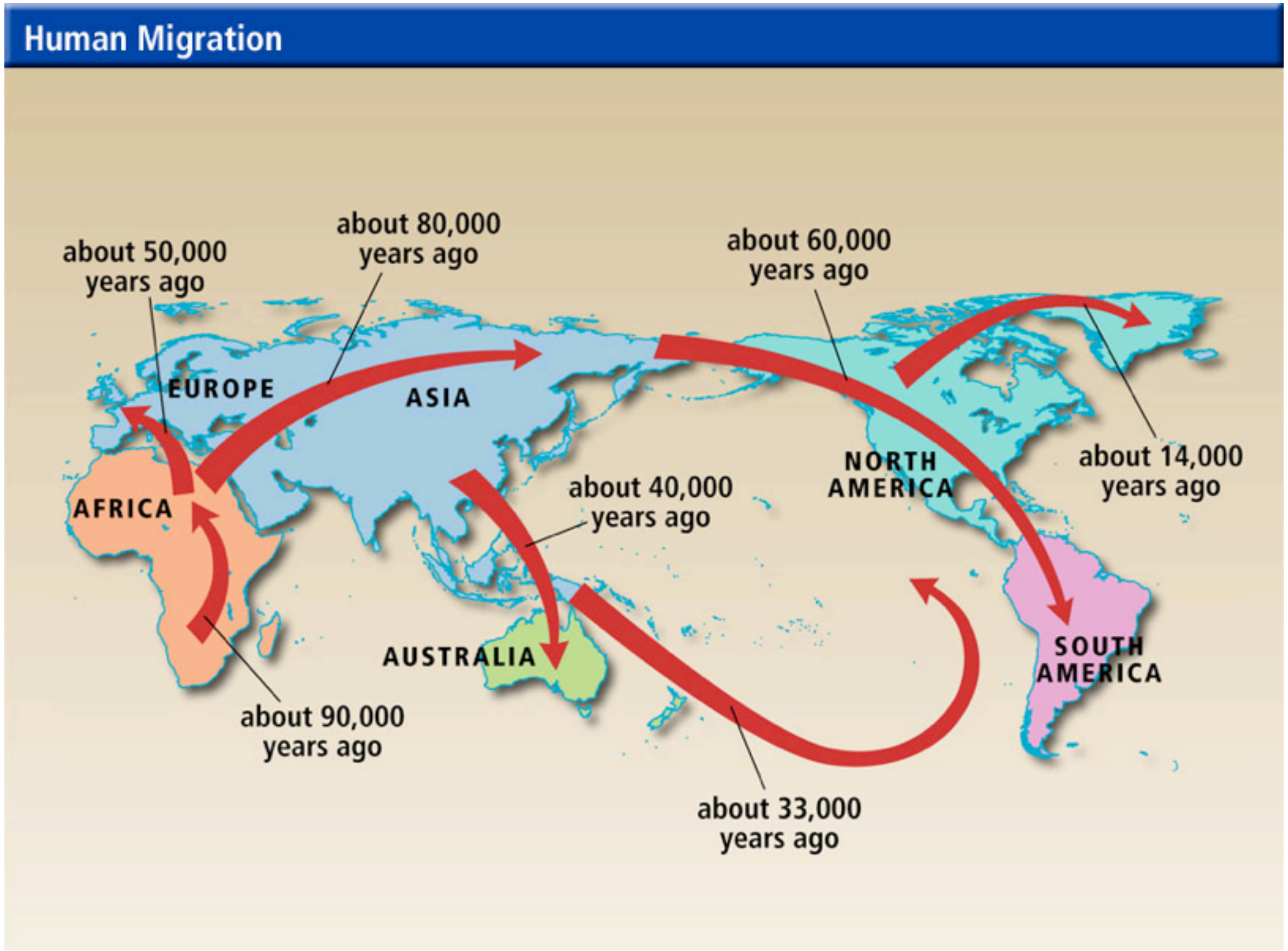
- Second, you will need to figure out which to include in your analysis (again, fuzzy rules!) but a good rule is if the parameter estimate associated with the covariate is large (=significant individual p-value) you should include it!

- 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)
  - <u>If these populations differ in frequencies of genotypes at a subset of measured genotypes / polymorphisms</u>
  - <u>If these populations differ in the mean value of the phenotype</u>
- 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:

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

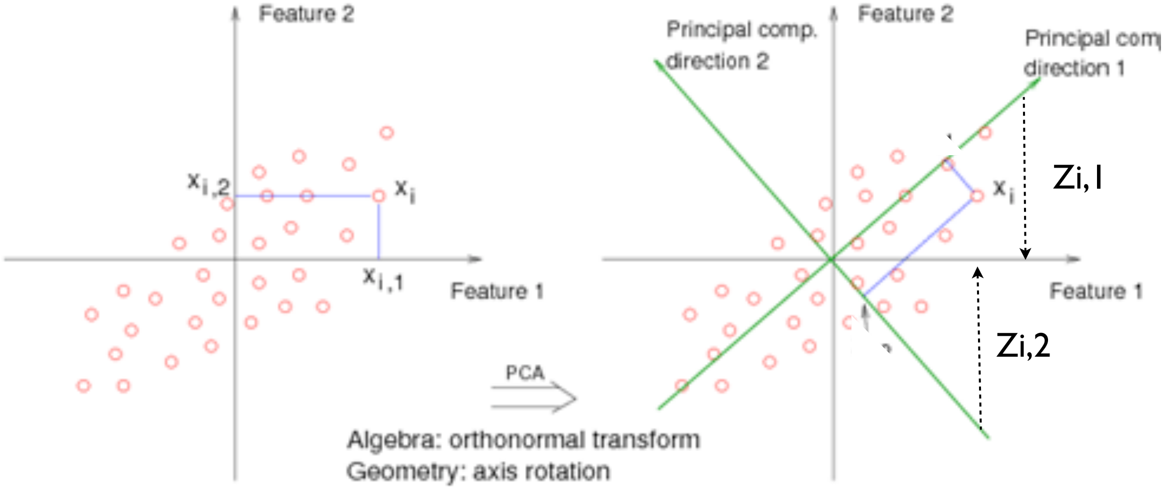
- How do we tell if our covariate correction "worked" well enough that we should interpret the results of our analysis?

Learning unmeasured population factors

- To learn a population factor, analyze the genotype data

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

- 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 $X_{Z,1} = Z_1$, $X_{Z,2} = Z_2$

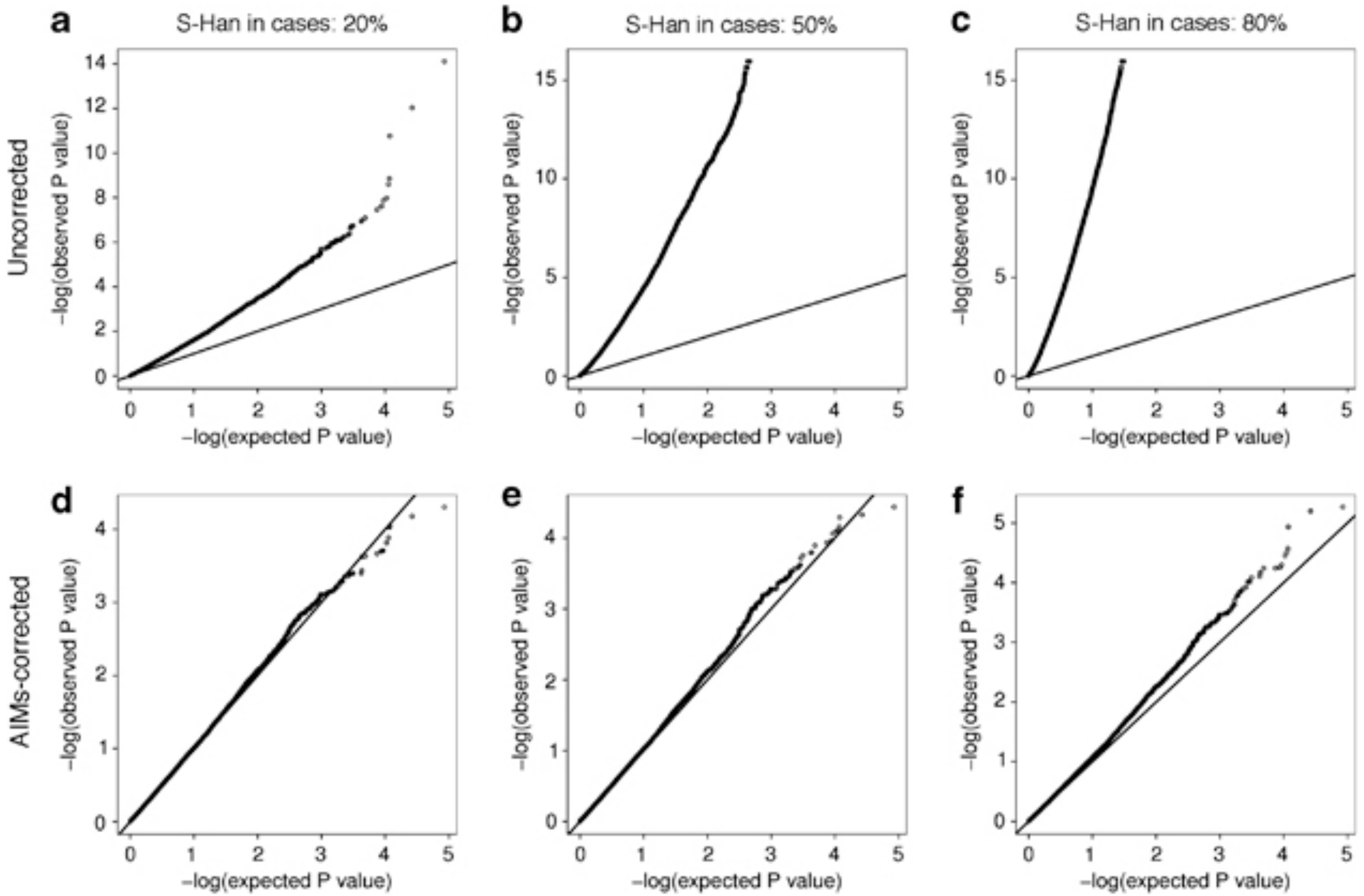
- You could also determine which individual is in which pop and define random variables for pop assignment, e.g. for two populations include single covariate by setting, $X_{Z,1}(\text{pop1}) = 1$, $X_{Z,1}(\text{pop2}) = 0$ (generally less optimal but can be used!)

- 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?)

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



That’s it for today

- Next lecture we will discuss minimum / minimal GWAS and begin our discussion of logistic regression!