Lecture 12: Introduction to genome-wide association studies (GWAS)

Jason Mezey
jgm45@cornell.edu
March 15, 2016 (T) 8:40-9:55
Announcements

• Homework #4 has been graded and available after class today

• Key for homework #5 will be posted today

• REMINDER: NO CLASS (OR COMPUTER LABS) THE FOLLOWING TWO WEEKS: March 22-March 31

• Midterm will be available tonight (!!): Due 11:59PM, Fri., March 18th

• REMINDER: we WILL have class Thurs., March 17, but NO computer lab this week (and no office hours)
Announcements II

Quantitative Genomics and Genetics - Spring 2016
BTRY 4830/6830; PBSB 5201.01

Midterm - available online 11:59PM, Tues., March 15

For midterm exam, due before 11:59PM, Fri., March 18

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 YOU ARE NOT ALLOWED TO COMMUNICATE OR IN ANY WAY ASK ANYONE FOR ASSISTANCE WITH THIS EXAM IN ANY FORM** (the only exceptions are Mahya, Jin, and Dr. Mezey). 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. A complete answer to this exam will include two files: a SINGLE text file including all of your R code, and a SINGLE file including all of your typed answers and plots (where the latter may be a scan as long as we can read it). Please note that for your R code, to get full credit for all problems, we must be able to run your code and replicate all of your results (with ease!). We will attempt to run your code if you do not do this but we will deduct points accordingly (note that no code = no credit!).

3. 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 you advantage to attempt every part of every question.

4. The exam must be in Mahya’s or Jin’s (as appropriate) email inbox before 11:59PM Fri., March 18. It is your responsibility to make sure that it is in the appropriate email box before then and no excuses will be accepted (power outages, computer problems, Cornell’s internet slowed to a crawl, etc.). Remember: you are welcome to hand this in early! We will deduct points for being late for exams received after this deadline (even if it is by minutes!!).

Good Luck!!
Summary of lecture 12

• Last lecture, we introduced *hypothesis testing* for the quantitative genetic model

• Today we will provide a rigorous introduction to the statistical foundation of Genome-wide Association Study (GWAS) analysis

• We will also discuss concepts in population genetics as these relate to linkage disequilibrium, a concept which in turn is critical for understanding analysis issues GWAS
Conceptual Overview

Genetic System

Does A1 → A2 affect Y?

Sample or experimental pop

Measured individuals (genotype, phenotype)

Regression model

Regress / DNR

Model params F-test

Pr(Y|X)
Review: genetic system

• Our goal in quantitative genetics / genomics is to identify loci (positions in the genome) that contain causal mutations / polymorphisms / alleles

• causal mutation or polymorphism - a position in the genome where an experimental manipulation of the DNA produces an effect on the phenotype on average or under specified conditions

• Formally, we may represent this as follows:

\[ A_1 \rightarrow A_2 \Rightarrow \Delta \overline{Y} \]

• Our experiment will be a statistical experiment (sample and inference!)
Review: the statistical model

- As with any statistical experiment, we need to begin by defining our sample space
- In the most general sense, our sample space is:
  \[
  \Omega = \{ \text{Possible Individuals} \}
  \]
- More specifically, each individual in our sample space can be quantified as a pair of sample outcomes so our sample space can be written as:
  \[
  \Omega = \{ \Omega_g \cap \Omega_P \}
  \]
  Where \( \Omega_g \) is the genotype sample space at a locus and \( \Omega_P \) is the phenotype sample space
- Note that genotype \( g_i = A_j A_k \) is the set of possible genotypes, where for a diploid, with two alleles:
  \[
  \Omega_g = \{ A_1 A_1, A_1 A_2, A_2 A_2 \}
  \]
- For the phenotype, this can be any type of measurement (e.g. sick or healthy, height, etc.)
Review: the statistical model

- Next, we need to define the probability model on the sigma algebra of the sample space ($\mathcal{F}_{\{g,P\}}$):
  \[ Pr(\mathcal{F}_{\{g,P\}}) \]

- Which defines the probability of each possible genotype and phenotype pair:
  \[ Pr\{g, P\} \]

- We will define two (types) or random variables ($\ast = \text{state does not matter}$):
  \[ Y : (\ast, \Omega_P) \to \mathbb{R} \]
  \[ X : (\Omega_g, \ast) \to \mathbb{R} \]

- Note that the probability model induces a (joint) probability distribution on this random vector (these random variables):
  \[ Pr(Y, X) \]
Review: genetic probability model

- The quantitative genetic model is a multiple regression model with the following independent (“dummy”) variables:

\[ X_a(A_1A_1) = -1, \quad X_a(A_1A_2) = 0, \quad X_a(A_2A_2) = 1 \]

\[ X_d(A_1A_1) = -1, \quad X_d(A_1A_2) = 1, \quad X_d(A_2A_2) = -1 \]

<table>
<thead>
<tr>
<th></th>
<th>( A_1A_1 )</th>
<th>( A_1A_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( A_2A_2 )</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>

- and the following regression equation:

\[ Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon \]

\[ \epsilon \sim N(0, \sigma_\epsilon^2) \]
Review: genetic inference

- For our model focusing on one locus:
  \[ Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon \]
  \[ \epsilon \sim N(0, \sigma_\epsilon^2) \]
- We have four possible parameters we could estimate:
  \[ \theta = [\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2] \]
- However, for our purposes, we are only interested in the genetic parameters and testing the following null hypothesis:
  \[ H_0 : Cov(X_a, Y) = 0 \cap Cov(X_d, Y) = 0 \]
  \[ H_A : Cov(X_a, Y) \neq 0 \cup Cov(X_d, Y) \neq 0 \]

\[ \text{OR} \]

\[ H_0 : \beta_a = 0 \cap \beta_d = 0 \]
\[ H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \]
Review: genetic estimation

- We will define a MLE for our parameters:
  \[ \beta = [\beta_\mu, \beta_a, \beta_d] \]

- Recall that an MLE is simply a statistic (a function that takes a sample in and outputs a number that is our estimate)

- In this case, our statistic will be a vector valued function that takes in the vectors that represent our sample
  \[ T(y, x_a, x_d) = \hat{\beta} = [\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d] \]

- Note that we calculate an MLE for this case just as we would any case (we use the likelihood of the fixed sample where we identify the parameter values that maximize this function)

- In the linear regression case (just as with normal parameters) this has a closed form:
  \[ MLE(\hat{\beta}) = (x^T x)^{-1} x^T y \]
Review: genetic hypothesis testing

- We are going to test the following hypothesis:
  \[ H_0 : \beta_a = 0 \cap \beta_d = 0 \]
  \[ H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \]

- To do this, we need to construct the following test statistic (for which we know the distribution!):
  \[ T(y, x_a, x_d| H_0 : \beta_a = 0 \cap \beta_d = 0) \]

- Specifically, we are going to construct a likelihood ratio test (LRT)

- This is calculated using the same structure that we have discussed (i.e. ratio of likelihoods that take values of parameters maximized under the null and alternative hypothesis)

- In the case of a regression (not all cases!) we can write the form of the LRT for our null in an alternative (but equivalent!) form

- In addition, our LRT has an exact distribution for all sample sizes \( n \) (!!)
Review: genetic inference

- We now have everything we need to construct a hypothesis test for:
  \[ H_0 : \beta_a = 0 \cap \beta_d = 0 \]
  \[ H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \]

- This is equivalent to testing the following:
  \[ H_0 : Cov(X, Y) = 0 \]

- For a linear regression, we use the F-statistic for our sample:
  \[ F_{[2, n-3]}(y, x_a, x_d) = \frac{MSM}{MSE} \]

- We then determine a p-value using the distribution of the F-statistic under the null:
  \[ pval(F_{[2, n-3]}(y, x_a, x_d)) \]
### Review: Genetic Inference

- To construct our LRT for our null, we will need several components, first the predicted value of the phenotype for each individual:

\[
\hat{y}_i = \beta_\mu + x_{i,a}\hat{\beta}_a + x_{i,d}\hat{\beta}_d
\]

- Second, we need the “Sum of Squares of the Model” (SSM) and the “Sum of Squares of the Error” (SSE):

\[
SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]

- Third, we need the “Mean Squared Model” (MSM) and the “Mean Square Error” (MSE) with degrees of freedom (df) \(df(M) = 3 - 1 = 2\) and \(df(E) = n - 3\):

\[
MSM = \frac{SSM}{df(M)} = \frac{SSM}{2} \quad MSE = \frac{SSE}{df(E)} = \frac{SSE}{n - 3}
\]

- Finally, we calculate our (LRT!) statistic, the F-statistic with degrees of freedom [2, n-3]:

\[
F_{[2,n-3]} = \frac{MSM}{MSE}
\]
Side-topic: Alternative (ANOVA) formulation

- Note that we can construct an equivalent formulation to our *linear regression* using an ANOVA coding.

- ANOVA stands for *ANalysis Of VAriance* and, despite the name, it is really a test of whether “means” of groups are different.

- A genetic ANOVA model is the same as our linear regression, except the “dummy” variables are coded differently (everything else is the same!)
Side-topic: Alternative (ANOVA) formulation II

- Remember the independent (dummy) variable coding for a regression is:
\[
X_a(A_1A_1) = -1, X_a(A_1A_2) = 0, X_a(A_2A_2) = 1
\]
\[
X_d(A_1A_1) = -1, X_d(A_1A_2) = 1, X_d(A_2A_2) = -1
\]

- The ANOVA coding is the following:
\[
X_{A_1A_1} = 1, X_{A_1A_2} = 0, X_{A_2A_2} = 0
\]
\[
X_{A_1A_1} = 0, X_{A_1A_2} = 1, X_{A_2A_2} = 0
\]
\[
X_{A_1A_1} = 0, X_{A_1A_2} = 0, X_{A_2A_2} = 1
\]

- The models corresponding to a linear regression and ANOVA are:
\[
Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon
\]
\[
Y = X_{A_1A_1}\beta_{A_1A_1} + X_{A_1A_2}\beta_{A_1A_2} + X_{A_2A_2}\beta_{A_2A_2} + \epsilon
\]
Side-topic: Alternative (ANOVA) formulation III

• For the ANOVA formulation, the parameters are:

\[ \theta = [\beta_{A_1A_1}, \beta_{A_1A_2}, \beta_{A_2A_2}] \]

• And we test the null hypothesis:

\[ H_0 : \beta_{A_1A_1} = \beta_{A_1A_2} = \beta_{A_2A_2} \]

\[ H_A : \beta_{A_jA_k} = \beta_{A_lA_m} \quad jk \neq lm \]

• Note that estimation (MLE) and the hypothesis test (F-test) construction are the same (=same equations)!!

• Why would we use an ANOVA formulation (what is the difference)?
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)
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!)
Solution: 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 causal genotype (!!)

- This is the case in genetic systems (why!?)

- Do we know which genotype is causal in this scenario?

\[
\begin{bmatrix}
1 & 0.96 & 0.83 \\
0.96 & 1 & 0.89 \\
0.83 & 0.89 & 1
\end{bmatrix}
\]
Solution: 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?
The Manhattan plot I

• We will consider a number of visualization tools for analyzing GWAS data

• For the moment, we will introduce the *Manhattan plot*

• This is a plot of genotypes on the x-axis and on the y-axis the \(-\log p\)-values (base 10) (why?) resulting from each hypothesis test of each genotype

• Each “point” on the plot is therefore a single \(p\)-value corresponding to a single measured genotype

• We are looking for sets of points with high \(-\log p\)-value = the position of a causal polymorphism
The Manhattan plot II: examples

MTRR

Chromosome

-Log\(P\)

-\(\log_{10}(P)\)

Pi-ta

-Log\(10\) (P)
Linkage Disequilibrium (LD)

- 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 disequilibrium includes both linkage disequilibrium AND other types of disequilibrium (!!), e.g. gametic phase disequilibrium.
Rigorous formulation of GWAS analysis I

- For a GWAS, we assume that there could be causal polymorphisms $X=(X_a, X_d)$ that are BOTH in the same physical position of the genome AND are correlated (=in linkage disequilibrium) with polymorphisms that we have measured $X'=(X_a', X_d')$:

$$|\text{Corr}(X, X')| >> 0$$

- Note we are using $|\text{Corr}(X,X')|$ and not specifying $X_a', X_d'$ because either or both $X_a', X_d'$ could be correlated with $X_a, X_d$

- For analysis of a GWAS with $N$ measured genotypes (2 alleles each) and a normal (error) phenotype we perform $N$ hypothesis tests:

$$H_0 : \beta'_a = 0 \cap \beta'_d = 0$$
$$H_A : \beta'_a \neq 0 \cup \beta'_d \neq 0$$

$$Y = \beta_\mu + X'_a \beta'_a + X'_d \beta'_d + \epsilon$$

- For genotypes / sets of genotypes (which sets?) for which we reject the null, we assume that this indicates a position of causal polymorphism (we have mapped the position)
Rigorous formulation of GWAS analysis II

- Polymorphisms for which we reject the null are “tags” of the (=in linkage disequilibrium with) causal polymorphisms

- The true parameter values of the regression model for the tags:
  \[ \beta'_a = 0 \quad \beta'_d = 0 \]

- The more correlated the tag with the causal polymorphism, the closer the tag parameter estimates to the parameter of the causal polymorphism (on average):
  \[ |\text{Corr}(X, X')| = 0 \quad |\text{Corr}(X, X')| \gg 0 \quad |\text{Corr}(X, X')| = 1 \]

  \[ \hat{\beta}'_a = 0, \hat{\beta}'_d = 0 \quad \hat{\beta}'_a = \beta_a, \hat{\beta}'_d = \beta_d \]

- This means that for tag markers, we are getting a really bad estimate of the true tag parameters but a “good” estimate of the causal polymorphism parameters, which is what we depend on in a GWAS (!!)
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 disequilibrium includes both linkage disequilibrium AND other types of disequilibrium (!!), e.g. gametic phase disequilibrium.

**Linkage Disequilibrium (LD)**

- LD
- equilibrium, linkage
Population versus quantitative genetics

- The theoretical explanations and statistical modeling used to understand linkage disequilibrium (LD) belong to the field of population genetics.

- **Population genetics (genomics)** - a field concerned with modeling the processes that lead to observed genomic variation in a population (i.e. the field is concerned with explaining patterns of DNA in a population).

- **Quantitative genetics (genomics)** - a field concerned with understanding and inferring the relationship between genotypes and phenotypes.

- Since LD describes the pattern of genomes in a population, understanding LD is the province of population genetics.

- We will discuss just the most basic population genetic concepts critical for our purposes but I encourage you to take a class in pop gen.
Hardy-Weinberg equilibrium I

• LD has two components: *linkage* and *(dis)*equilibrium

• Linkage refers to *physical linkage* of alleles on a chromosome, e.g. if alleles A1 at one locus and B1 at a second locus are on the same chromosome (they are on the same molecule) they are linked

• Disequilibrium refers to any set of alleles at two or more loci that are not in Hardy-Weinberg (H-W) equilibrium

• H-W equilibrium to a *statistical description* of the pattern of alleles in a population

• If a population is in H-W equilibrium: 1. alleles of an individual polymorphism are in equilibrium and 2. sets of alleles at two or more polymorphic sites are in equilibrium (where the later concerns LD)
Hardy-Weinberg equilibrium II

- Theoretical conditions for that lead to H-W equilibrium are defined within population genetics, i.e. infinite population size, random mating, no selection, no mutation, no gene flow, meiotic drive, or linkage (do these ever apply in real populations!?)

- 1. Under H-W equilibrium, if we represent the probability of the A1 allele in a population or a sample as $Pr(A1)=p$ and the probability of the A2 allele as $Pr(A2)=q=1-p$, then the probability of the genotypes in the population are:

\[ Pr(A_1, A_1) = Pr(A_1)Pr(A_1) = p^2 \]
\[ Pr(A_1, A_2) = 2Pr(A_1)Pr(A_2) = 2pq \]
\[ Pr(A_2, A_2) = Pr(A_2)Pr(A_2) = q^2 \]

- 2. Under H-W equilibrium, for two polymorphic sites A and B, for all i, j, k, and l we have:

\[ Pr(A_i A_j B_k B_l) = Pr(A_i A_j)Pr(B_k B_l) = Pr(A_i)Pr(A_j)Pr(B_k)Pr(B_l) \]
\[ \Rightarrow (Corr(X_{a,A}, X_{a,B}) = 0) \cap (Corr(X_{a,A}, X_{d,B}) = 0) \]
\[ \cap (Corr(X_{d,A}, X_{a,B}) = 0) \cap (Corr(X_{d,A}, X_{d,B}) = 0) \]
Hardy-Weinberg equilibrium III

- Note that H-W refers to independence of alleles occurring together both WITHIN a polymorphic site and BETWEEN polymorphic sites:

\[ Pr(A_iA_jB_kB_l) = Pr(A_iA_j)Pr(B_kB_l) = Pr(A_i)Pr(A_j)Pr(B_k)Pr(B_l) \]

- Notation note 1: Instead of probabilities of alleles and genotypes, we often refer to these as frequencies (for our purposes, they are equivalent!)

- Notation note 2: We have some notation gridlock in the literature, e.g. “p” for \( Pr(A_1) \) (also a parameter of binomial!!) and \( x_{ij} \) is often used to refer to the probability / frequency of \( A_iA_j \) (instead of as a r.v.!!)

- For a sample of \( n \) individuals in a sample or population, the less frequent allele (=the allele with the lower probability) is the minor allele and this frequency (=probability) is the minor allele frequency (MAF)

- As an example, consider a sample of \( n=5 \) individuals with the following observed genotypes: \( A1A1, A1A2, A1A2, A2A2, A2A2 \), then the frequency of allele \( A1 = 4/10 \), the frequency of allele \( A2 = 6/10 \) and allele \( A1 \) is the minor allele and the MAF = 0.4
Linkage Disequilibrium (LD) I

- Two polymorphic sites in the genome are in Disequilibrium if:
  \[ Pr(A_i B_j, A_k B_l) \neq Pr(A_i A_k) Pr(B_j B_l) \]
  \[ \Rightarrow (Corr(X_{aA}, X_{aB}) \neq 0) \cup (Corr(X_{aA}, X_{dB}) \neq 0) \]
  \[ \cup (Corr(X_{dA}, X_{aB}) \neq 0) \cup (Corr(X_{dA}, X_{dB}) \neq 0) \]

- Two polymorphic sites in the genome in Linkage Disequilibrium (LD) if they are in Disequilibrium AND they physically linked on a chromosome (!!!)

- Note that Disequilibrium (and LD) do not refer to H-W within a polymorphic site ONLY between sites, e.g. two sites may be in LD but the genotype frequencies within the two sites may be in H-W equilibrium

- We now know what the correlation between tag and causal polymorphism is referring to (!!!), simply set one to A (=X') and the other to B (=X)!
Linkage Disequilibrium (LD) II

- Note that the value of LD for identifying causal polymorphisms depends on the following:
  - With a tag depends on there not being disequilibrium among polymorphisms that are not physically linked (!!!)
  - Polymorphisms that are close to each tend to be in higher LD than those that are further apart on a chromosome
- Let’s next consider the biological explanation as to why this is the case, starting with polymorphisms on different chromosomes and then with polymorphisms on the same chromosome
That's it for today

- Next lecture: continued discussion of GWAS statistical, analysis, and interpretation issues!