Lecture 13: Intro to GWAS and population genetic, statistical, and experimental issues

Jason Mezey
jgm45@cornell.edu
March 14, 2019 (Th) 10:10-11:25
Announcements

• Reminder: office hours tomorrow (Fri., March 15), 2-4PM and on Mon. (March 18), 2-4PM
Summary of lecture 13

• Last lecture, we provided a rigorous introduction to hypothesis testing in genetics

• Today we will introduce Genome-wide Association Study (GWAS) analysis (!!)

• We will also discuss important concepts in population genetics and discuss statistical / experimental issues that impact GWAS analysis and interpretation
Conceptual Overview

Genetic System

Does A1 $\rightarrow$ A2 affect Y?

Sample or experimental pop

Measured individuals (genotype, phenotype)

Regressions model

Pr(Y|X)

Reject / DNR

Model params F-test
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:

\[
A_1 \rightarrow A_2 \Rightarrow \Delta Y | Z
\]

- Our experiment will be a statistical experiment (sample and inference!)
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:
  \[ H_0 : \text{Cov}(X_a, Y) = 0 \cap \text{Cov}(X_d, Y) = 0 \]
  OR
  \[ H_0 : \beta_a = 0 \cap \beta_d = 0 \]

  \[ H_A : \text{Cov}(X_a, Y) \neq 0 \cup \text{Cov}(X_d, Y) \neq 0 \]
  OR
  \[ H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \]
Review: genetic inference II

- Recall that inference (whether estimation or hypothesis testing) starts by collecting a sample and defining a statistic on that sample.

- In this case, we are going to collect a sample of $n$ individuals where for each we will measure their phenotype and their genotype (i.e. at the locus we are focusing on).

- That is an individual $i$ will have phenotype $y_i$ and genotype $g_i = A_jA_k$ (where we translate these into $x_a$ and $x_d$).

- Using the phenotype and genotype we will construct both an estimator (a statistic!) and we will additionally construct a test statistic.

- Remember that our regression probability model defines a sampling distribution on our sample and therefore on our estimator and test statistic (!!!).
Review: genetic estimation

- Let's look at the structure of this estimator:

\[ y = x\beta + \epsilon \]

\[
\begin{bmatrix}
    y_1 \\
    y_2 \\
    \vdots \\
    y_n
\end{bmatrix}
= \begin{bmatrix}
    1 & x_{1,a} & x_{1,d} \\
    1 & x_{2,a} & x_{2,d} \\
    \vdots & \vdots & \vdots \\
    1 & x_{n,a} & x_{n,d}
\end{bmatrix}
\begin{bmatrix}
    \beta_\mu \\
    \beta_a \\
    \beta_d
\end{bmatrix}
+ \begin{bmatrix}
    \epsilon_1 \\
    \epsilon_2 \\
    \vdots \\
    \epsilon_n
\end{bmatrix}
\]

\[
MLE(\hat{\beta}) = (x^T x)^{-1} x^T y
\]

\[
MLE(\hat{\beta}) = \begin{bmatrix}
    \hat{\beta}_\mu \\
    \hat{\beta}_a \\
    \hat{\beta}_d
\end{bmatrix}
\]
Review: genetic hypothesis test I

• 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 hypothesis test II

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

  $$\hat{y}_i = \hat{\beta}_\mu + x_{i,a}\hat{\beta}_a + x_{i,d}\hat{\beta}_d$$

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

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

- Third, we need the “Mean Squared Model” (MSM) and the “Mean Square Error” (MSE) with degrees of freedom (df)

  $$df(M) = 3 - 1 = 2 \text{ and } \quad df(E) = n - 3$$

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

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

  $$F_{[2,n-3]} = \frac{MSM}{MSE}$$
In general, the F-distribution (continuous random variable!) under the H0 has variable forms that depend on d.f.:

Note when calculating a p-value for the genetic model, we consider the value of the F-statistic we observe or more extreme towards positive infinite (!!) using the F-distribution with \([2, n=3]\) d.f.

However, also this is actually a two-tailed test (what is going on here (!?))
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)
• 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!)
Review: 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?
Review: 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 disequilibrium includes both linkage disequilibrium AND other types of disequilibrium (!!), e.g. gametic phase disequilibrium.
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
The Manhattan plot: examples

MTRR

Chromosome

-Log(P)

-log_{10}(P)

-Log_{10}(P)

Pi–ta

Chromosome

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

1 2 3 4 5 6 7 8 9 10 11 12

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
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.
Different chromosomes I

- Polymorphisms on different chromosomes tend to be in equilibrium because of independent assortment and random mating, i.e. random matching of gametes to form zygotes.
Different chromosomes II

- Polymorphisms on different chromosomes tend to be in equilibrium because of independent assortment and random mating, i.e. random matching of gametes to form zygotes.
Different chromosomes III

• More formally, we represent independent assortment as:

\[ Pr(A_iB_k) = Pr(A_i)Pr(B_k) \]

• For random pairing of gametes to produce zygotes:

\[ Pr(A_iB_k, A_jB_l) = Pr(A_iB_k)Pr(A_jB_l) \]

• Putting this together for random pairing of gametes to produce zygotes we get the conditions for equilibrium:

\[ Pr(A_iB_k, A_jB_l) = Pr(A_iB_k)Pr(A_jB_l) \]

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

\[ \Rightarrow (Corr(X_{a,A}, X_{a,B}) = 0) \cap (Corr(X_{a,A}, X_{d,B}) = 0) \]

\[ \cap (Corr(X_{d,A}, X_{a,B}) = 0) \cap (Corr(X_{d,A}, X_{d,B}) = 0) \]
Same chromosome I

- For polymorphisms on the same chromosome, they are linked so if they are in disequilibrium, they are in LD

- In general, polymorphisms that are closer together on a chromosome are in greater LD than polymorphisms that are further apart (exactly what we need for GWAS!)

- This is because of recombination, the biological process by which chromosomes exchange sections during meiosis

- Since recombination events occur at random throughout a chromosome (approximately!), the further apart two polymorphisms are, the greater the probability of a recombination event between them

- Since the more recombination events that occur between polymorphisms, the closer they get to equilibrium, this means markers closer together tend to be in greater LD
• In diploids, recombination occurs between pairs of chromosomes during meiosis (the formation of gametes)

• Note that this results in taking alleles that were physically linked on different chromosomes and physically linking them on the same chromosome
Same chromosome III

- To see how recombination events tend to increase equilibrium, consider an extreme example where alleles A1 and B1 always occur together on a chromosome and A2 and B2 always occur together on a chromosome:

\[ Pr(A_1B_2) = 0, \ Pr(A_2B_1) = 0 \]

\[ Corr(X_{a,A}, X_{a,B}) = 1 \text{ AND } Corr(X_{d,A}, X_{d,B}) = 1 \]

- If there is a recombination event, most chromosomes are A1-B1 and A2-B2 but now there is an A1-B2 and A2-B1 chromosome such that:

\[ Pr(A_1B_2) \neq 0, \ Pr(A_2B_1) \neq 0 \]

\[ Corr(X_{a,A}, X_{a,B}) \neq 1 \text{ AND } Corr(X_{d,A}, X_{d,B}) \neq 1 \]

- Note recombination events disproportionally lower the probabilities of the more frequent pairs!

- This means over time, the polymorphisms will tend to increase equilibrium (decrease LD)

- Since the more recombination events, the greater the equilibrium, polymorphisms that are further apart will tend to be in greater equilibrium, those closer together in greater LD
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.

```
Chr. 1
A   B   C

Chr. 2
D
```

LD  
\[\text{equilibrium, linkage}\]

\[\text{equilibrium, no linkage}\]
Side topic: connection coin flip models to allele / genotypes

- Recall we the one coin flip example (how does the parameter of Bernoulli relate to MAF?):
  \[ \Omega = \{ H, T \} \quad X(H) = 0, X(T) = 1 \]
  \[ Pr(X = x|p) = P_X(x|p) = p^x(1-p)^{1-x} \]

- The following model for two coin flips maps perfectly on to the model of genotypes (e.g., represented as number of A1 alleles) under Hardy-Weinberg equilibrium (e.g., for MAF = 0.5):
  \[ X(HH) = 0, X(HT) = 1, X(TH) = 1, X(TT) = 2 \]
  \[ Pr(HH) = Pr(HT) = Pr(TH) = Pr(TT) = 0.25 \]

  \[ P_X(x) = Pr(X = x) = \begin{cases} 
  Pr(X = 0) = 0.25 \\
  Pr(X = 1) = 0.5 \\
  Pr(X = 2) = 0.25 
\end{cases} \]

  \[ Pr(X = x|n, p) = P_X(x|n, p) = \binom{n}{x} p^x(1-p)^{n-x} \]

- Note that the model need not conform to H-W since consider the following model (we could use a multinomial probability distribution):

  \[ Pr(X_1 = 0, X_2 = 0) = 0.0, Pr(X_1 = 0, X_2 = 1) = 0.25 \]
  \[ Pr(X_1 = 1, X_2 = 0) = 0.25, Pr(X_1 = 1, X_2 = 1) = 0.25 \]
  \[ Pr(X_1 = 2, X_2 = 0) = 0.25, Pr(X_1 = 2, X_2 = 1) = 0.0 \]
Patterns and 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:)

![Image of LD patterns and representing LD]
Measuring LD I

- There are *many* statistics used to represent LD but we will present the two most common.

- For the first, define the correlation:

\[
r = \frac{\Pr(A_i, B_k) - \Pr(A_i) \Pr(B_k)}{\sqrt{\Pr(A_i)(1 - \Pr(A_i))} \sqrt{\Pr(B_k)(1 - \Pr(B_k))}}
\]

- As a measure of LD, we will consider this squared:

\[
r^2 = \frac{(\Pr(A_i, B_k) - \Pr(A_i) \Pr(B_k))^2}{(\Pr(A_i)(1 - \Pr(A_i)))(\Pr(B_k)(1 - \Pr(B_k)))}
\]

- Note that this is always between one and zero!
Measuring LD II

- A “problem” with r-squared is that when the MAF of A or B is small, this statistic is small

- For the second measure of LD, we will define a measure $D'$ that is not as dependent on MAF:

\[
D = Pr(A_i, B_k) - Pr(A_i)Pr(B_k)
\]

\[
D' = \frac{D}{\min(Pr(A_1B_2), Pr(A_2, B_1))} \text{ if } D > 0
\]

\[
D' = \frac{D}{\min(Pr(A_1B_1), Pr(A_2, B_2))} \text{ if } D < 0
\]

- Note that this is always between -1 and 1 (!!)
That’s it for today

• See you Tues.!