Quantitative Genomics and Genetics
BTRY 4830/6830; PBSB.5201.03

Lecture 26: Inbred line analysis and Evolutionary Quantitative Genomics

Jason Mezey
jgm45@cornell.edu
May 10, 2022 (T) 8:05-9:20AM
Announcements

- Last computer labs this week (Optional)!
- Last lecture this coming Tues. (May 10), also optional!
- REMINDER: your PROJECT is due today (May 10) by 11:59PM (!!)
- REMINDER: you FINAL (take home, open book, NO DISCUSSION of any type with anyone about the final once the exam has started) will be available tomorrow (May 11) and will be due May 21
  - We are giving you 10 days to complete the final, but (if you are prepared) it will only take you a few hours to complete
  - SAME format (and rules!) as the midterm
  - You will need to perform both a linear AND logistic regression (using the equations / algorithms introduced in class) both with and without covariates (maybe more than 1 covariate!)
Final Exam

Available on CMS on Weds., May 11
Due 11:59PM (ET) Sat., May 21

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 e.g., DO NOT POST PUBLIC MESSAGES ON PIAZZA! (the only exceptions are Beulah, Yajas, 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 in Rmarkdown, where you will submit your .Rmd script and associated .pdf file. 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!!). You should include all of your plots and written answers in this same .Rmd script with your R code.

4. The exam must be uploaded on CMS before 11:59PM (ET) Thurs., 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!!).
Summary of Lecture 26

Today we will complete our discussion of inbred line analysis (traditional agricultural genetics analysis) and evolutionary quantitative genetics analysis to complete a (brief) discussion of the foundations of Quantitative Genomics!
Analysis of inbred lines

• **inbred line design** - a sampling experiment where the individuals in the sample have a known relationship that is a consequence of controlled breeding

• Note that the relationships may be know exactly (e.g. all individuals have the same grandparents) or are known within a set of rules (e.g. the individuals were produced by brother-sister breeding for $k$ generations)

• Note that inbred line designs are a form of **pedigrees** (= a sample of individuals for which we have information on relationships among individuals)
Historical importance of inbred lines

• Inbred lines have played a critical role in agricultural genetics (actually, both inbred lines and pedigrees have been important)

• This is particularly true for crop species, where people have been producing inbred lines throughout history and (more recently) for the explicit purposes of genetic analysis

• In genetic analysis, these have played an important historical role, leading to the identification of some of the first causal polymorphisms for complex (non-Mendelian!) phenotypes
GWAS definitions III

- **Quantitative Trait Locus (QTL)** - a causal polymorphism (or the locus containing the polymorphism) OR a large section of the genome containing a causal (or several!) polymorphisms.

- **expression Quantitative Trait Locus (eQTL)** - a QTL for a gene expression phenotype, i.e. a quantitative measurement of transcription level of a gene in a tissue.

- **xQTL** - a QTL for a next-generation sequencing technology measured phenotype, e.g. methylation, CHiP-Seq.

- **Quantitative trait nucleotide (QTN)** - a SNP that is a causal polymorphism (QTR for any polymorphism).
Importance of inbred lines

• Inbred lines continue to play a critical role in both agriculture (most plants we eat are inbred!) and in genetics

• The reason they continue to be important in genetic analysis is we can control the genetic background (e.g. epistasis!) and, once we know causal polymorphisms, we can integrate the section of genome containing the causal polymorphism through inbreeding designs or now through “exact” approaches like CRISPR (or TALEN) (!!)

• Where they used to be critically important in Quantitative Genetics was when we had access to many fewer genetic markers, inbreeding designs allowed “strong” inference for the markers in between

• This usage is less important now, but for understanding the Quant Gen literature (e.g. the specialized mapping methods applied to these line) we will consider several specialized designs and how we analyze them

• How should I analyze (high density) marker data for inbred lines? = do a GWAS analysis one market a time (!!) (maybe use a mixed model to account for inbred line structure…)
Types of inbred line designs (important in genetic analysis)

- A few main examples (non-exhaustive!):
  - B1 (Backcross) - cross between two inbred lines where offspring are crossed back to one or both parents
  - F2 - cross between two inbred lines where offspring are crossed to each other to produce the mapping population
  - NILs (Near Isogenic Lines) - cross between two inbred lines, followed by repeated backcrossing to one of the parent populations, followed by inbreeding
  - RILs (Recombinant Inbred Lines) an F2 cross followed by inbreeding of the offspring
  - Isofemale lines - offspring of a single female from an outbred (=non-inbred!) population are inbred

- We will discuss NILs and briefly mention the F2 design to provide a foundation for the major concepts in the literature
Consequences of inbreeding

• The reason that inbred line designs are useful is we can infer the unobserved markers (with low error!) even with very few markers.

• The reason is inbred lines designs result in homozygosity of the resulting lines (although they may be homozygous for different genotype!)

• Therefore, inbreeding, in combination with uncontrolled random sampling (=genetic drift) results in lines that are homozygous for one of the genotypes of the parents.
Example 1: NILs 1

Inbred line A (homozygous)  Inbred line B (homozygous)  Inbred line A (homozygous)  Backcross 1 (from 1st cross)  Inbred line A (homozygous)  Backcross 2 (from 2nd cross)

X

Additional backcrosses

Result:
Many lines that are homozygous, mostly (isogenic) red, each with a (different) blue homozygous regions (=near isogenic) etc.

Inbreeding of resulting offspring (after final backcross)
Example 1: NILs II

- For a “panel” (=NILs produced from the same design) since one marker allele from the “blue” lines within a blue region is to know the genotypes of the entirety of the region (i.e. it is from the blue lines), by individual marker testing, we can identify a polymorphism down to the size of the overlapping (“introgressed”) blue regions.

- e.g. for a marker indicated by the arrow where a regression model indicates the “blue” marker allele is associated with a larger phenotype on average than the “red” marker allele:
Example 2: interval mapping (F2)

- A limitation of NILs is the resolution is the size of the smallest “introgressed” region.

- The goal of “interval mapping” is to take advantage of different designs but with many possible recombination events, so we could map to a smaller region with a pedigree analysis approach.

- Recall the general structure of the pedigree likelihood equation (note we could also use a Bayesian approach!):

$$Pr(Y|X_{cp=Q})Pr(X_{cp}|X, r_{(X_{cp=Q}, X)}) = \sum_{\Theta_g} \prod_{i} Pr(y_i|g_i)Pr(g_i)Pr(g_i) \prod_{j=f+1}^{n} Pr(y_j|g_j)Pr(g_j, g_j, f, g_j, m, r)$$

- For interval mapping, we will use a version of this equation (what assumptions!??) to infer the state of unmeasured polymorphism “Q” that is in the proximity of markers we have measured:

$$Pr(Y|X_{cp=Q})Pr(X_{cp}|X, r_{(X_{cp=Q}, X)}) = \prod_{i} \sum_{\Theta_g} Pr(y_i|g_i, Q)Pr(g_i, Q|g_i, A, g_i, B, r)$$

- The first of these equations is just our glm (!!) or similar penetrance model, where we will consider an example of one type of inbreeding design (F2) to show the structure of the second
Example 2: interval mapping (F2)

Inbred line A (homozygous)  Inbred line B (homozygous)

X

F1 (cross these to each other)

F2
Example 2: interval mapping (F2)

F1 Gametes:

A_1 Q_1 B_1 A_2 Q_2 B_2

\frac{(1-r_{AQ})(1-r_{QB})}{2} \quad \frac{(1-r_{AQ})(1-r_{QB})}{2} \quad \frac{r_{AQ}(1-r_{QB})}{2} \quad \frac{r_{AQ}(1-r_{QB})}{2} \quad \frac{(1-r_{AQ})r_{QB}}{2} \quad \frac{(1-r_{AQ})r_{QB}}{2} \quad \frac{r_{AQ}r_{QB}}{2} \quad \frac{r_{AQ}r_{QB}}{2}
Example 2: interval mapping (F2)

\[
\Pr(\bar{Q}_1\bar{Q}_1|A_1A_1B_1B_1) = \frac{\Pr(A_1A_1\bar{Q}_1\bar{Q}_1B_1B_1)}{\Pr(A_1A_1B_1B_1)} = \frac{\left(\frac{(1-r_{AQ})(1-r_{QB})}{2}\right)^2}{\left(\frac{1-r_{AB}}{2}\right)^2} = \frac{(1-r_{AQ})^2(1-r_{QB})^2}{(1-r_{AB})^2}
\]
We can therefore substitute these conditional probabilities into our main equation and calculate the likelihood over possible values of $r$

$$Pr(Y | X_{cp=Q}) Pr(X_{cp} | X, r(X_{cp=Q}, X)) = \prod_{i} \sum_{\Theta_g} Pr(y_i | g_i, Q) Pr(g_i, Q | g_i, A, g_i, B, r)$$

In practice we perform a LRT comparing the null of no causal polymorphism for an alternative where there is a causal polymorphism in the marker defined region, where if we reject, we consider there to be a causal polymorphism in the region.

Note that the LRT is sometimes expressed as a “LOD” score (just LRT base 10!), which is just LRT times a constant (!!)

Note that once we have rejected the null for a region, we can identify the position within the interval by finding the position where a given value of $r$ maximizes the likelihood, i.e. hence “interval mapping”

We can translate this to a relative position if we have a physical map and recombination map (another complex subject!)

Example 2: interval mapping (F2) - see 2016
Value of interval mapping

- Similar to the case of using a linkage (pedigree) analysis to map causal polymorphisms for complex (non-Mendelian) phenotypes, in practice, interval mapping turns out to be not very useful.

- The reason is the same as in interval mapping (for complex phenotypes) that fitting a complex model does not provide very exact inferences.

- This is not to say inbred line designs are not useful (remember: the control of genetic background, etc.) but the best approach for analyzing these data is to test one marker at a time, i.e. just like in a GWAS!

- Given that we can now easily produce many markers across a region, we would get the same result as the ideal interval mapping result (!!)

- Interval mapping (and the many variants) is therefore (should) no longer used but understanding this technique is important for interpreting the literature (!!)
The last concepts we will discuss are from the field of genetics before we knew about DNA (!!) and therefore before genetic markers.

A way of thinking about the field of genetics before genetic markers was geneticists used the observation of the similarity between relatives to determine how much they could explain about underlying genetics (they could infer quite a bit!)

These inferences were used to model the patterns of phenotypes they observed in populations, how phenotypes evolved (=how the mean of a phenotype in a population changed over time), to guide plant and animal breeding to produce desired changes in phenotypes, etc.

The history goes back > 100 years where many of the concepts are important and continue to re-appear in quantitative genomics.
Intro. to classic quantitative genetics II

- We can understand the major concepts in classic quantitative genomics using our glm framework (!!)
- We will focus on phenotypes with normal error (= linear regression) but the concepts generalize
- The most important concept for understanding classic quantitative genetics is understanding narrow sense heritability (often just referred to as heritability), which is a property of a phenotype we measure:

\[ h^2 = \frac{V_A}{V_P} \]

- Note that this is a fraction with *additive genetic variance* (VA) in the numerator and *phenotypic variance* (VP) in the denominator
- The strange notation comes from a derivation by Sewall Wright (there are several derivations of heritability!) using path analysis, a type of probabilistic graphical model called a structural equation model
Why heritability is important

- RA Fisher used it to resolve the Mendelian versus Biometry argument that had gone on for ~30 years (with one paper!!) showing that a single genetic model could explain both patterns of inheritance

- RA Fisher also used heritability to demonstrate why Darwin’s evolution by natural selection was not only possible but occurred under extremely plausible conditions (“Fisher’s fundamental theorem”):

\[ \Delta \bar{w} = h_w^2 V_P \]

- More generally for evolution, heritability determines whether a phenotype changes under selection or genetic drift:

\[ \Delta \bar{Y} = h^2 s \]
\[ V_{\bar{P},t+1} = \frac{h_t^2 V_{P,t}}{N_e} \]

- We can use parts of heritability (additive genetic variance) to predict the relative offspring phenotype values from breeding two individuals (= breeding values)

- One of the most robust observations in biology: all reasonable phenotypes have non-zero heritability (!!), implying at least one causal polymorphism affects every phenotype (what else does it imply!!)
The components of heritability

- Recall that heritability is a fraction of two terms:
  \[ h^2 = \frac{V_A}{V_P} \]

- The denominator is the total variance for the phenotype (\(V_P\)), which we can calculate for the entire population as follows (or estimate using a sample):
  \[ V_P = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \]

- The numerator is the additive genetic variance (\(V_A\)) in the phenotype, which can be calculated for any phenotype (regardless of the complexity of the genetics!)

- However, this is easiest to understand when assuming there is a single causal polymorphism for the phenotype

- In this case, the VA is the following where the parameter is from our linear regression term only fitting the “additive” term (not dominance term!!):
  \[ V_A = 2MAF(1 - MAF)\beta_{\alpha}^2 = 2p(1 - p)\beta_{\alpha}^2 \]
Additive genetic variance I

- Recall that in our original regression (for a single causal polymorphism and assume we are fitting this model for the actual causal polymorphism, not a marker in LD!), we had two dummy variables and two parameters:

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

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

- For additive genetic variance, we will only define one dummy variable (even if there is dominance in the system!):

\[ X_\alpha(A_1A_1) = -1, X_\alpha(A_1A_2) = 0, X_\alpha(A_2A_2) = 1 \]

\[ Y = \beta_\mu + X_\alpha\beta_\alpha + \epsilon \]

- Given this model, it should be clear that the effects of dominance end up in the error term (!!!) just as for the case with un-modeled covariates

- We can then derive the additive genetic variance as follows:

\[ V_A = 2p(1-p)\beta_\alpha^2 \]
Additive genetic variance II

• There is a consequence of whether we fit two or one “slope” parameters in our regression model

• If we consider two slope parameters \( \beta_a, \beta_d \) (as we have done all semester!) the true values of the parameters are the same regardless of the allele frequency (MAF) of the causal polymorphism

• If we consider one regression parameter \( \beta_\alpha \) the true value of this parameter depends on the allele frequency (MAF) of the causal polymorphism

• The latter means that the true parameter value will change with changes in allele frequencies (!!)

• Stated another way, if we were to estimate this additive genetic regression parameter, there would be a different correct answer depending on the allele frequency in the population (!!)
Example how the parameter changes with MAF I

- Consider a case where there is dominance but we only fit the following model: \( Y = \beta_\mu + X_\alpha \beta_\alpha + \epsilon \)

MAF=0.5, larger \( \beta_\alpha \)

MAF=0.1, smaller \( \beta_\alpha \)

- Remember (!!) this is not the case if we fit two parameters: \( \beta_a, \beta_d \)
Example how the parameter changes with MAF I

- In a case of over-dominance (or under-dominance) with the right allele frequency, the true value of the parameter can be zero (!!):
In a purely additive case (no dominance) the parameter $\beta_\alpha$ does not change, regardless of MAF:

This makes sense since we only need the parameters $\beta_\mu, \beta_\alpha$ to completely fit the system.
Change in additive genetic variance with MAF

- Remember that additive genetic variance is a function of MAF:
  \[ V_A = 2MAF(1 - MAF)\beta_\alpha^2 = 2p(1 - p)\beta_\alpha^2 \]

- Additive genetic variance may therefore change (!!) with allele frequency, since the parameter \( \beta_\alpha \) may change

- The additive genetic variance is also a function of allele frequencies (MAF) so it may change due to allele frequencies through this term as well
Change in heritability with MAF

- Since additive genetic variance can change, it should be no surprise that heritability can change as well:

\[ h^2 = \frac{V_A}{V_P} = \frac{2p(1 - p)\beta^2}{V_P} \]

- Note that both the \( V_A \) and \( V_P \) can change with allele frequency since \( V_P \) includes the variance attributable to \( V_A \) (!!)

- Thus, heritability of a phenotype depends on the allele frequency in the population (!!)
Heritability concepts 1

• For multiple loci that are not in LD and when there is no epistasis, the additive genetic variance is:

\[ V_A = \sum_{i} 2p_i(1 - p_i)\beta_{\alpha,i}^2 \]

• The equations get more complex for LD and epistasis (and for more alleles, etc.)

• Note that even if the equations for \( V_A \) are complex for such cases, we can still estimate \( V_A \) for genetic systems (!!!)
Heritability concepts II

• We can estimate heritability using the resemblance between relatives, for example a parent-offspring regression (this was the origin of regression btw!)

• When regressing offspring phenotype values on the average value of their parents, the slope of the regression line is the heritability (under certain assumptions...) so an estimate of the slope is an estimate of heritability:

• There are many relationships that can be leveraged for this and the estimation procedures can involve many complex details (!!), e.g. pedigree analyses, mixed models, etc.
Heritability concepts III

- In agricultural genetics, we are often interested in value for an individual that reflects the value for which it will tend to increase or decrease the phenotype from the mean.

- e.g. if will breeding one bull to cows increase milk production compared to the results of breeding a different bull to these same cows?

- The breeding value (more specifically an estimate of the breeding value!) is used for this purpose, which we can derive from heritability (this concept requires more time than we have here).
Heritability concepts IV

- In classic quantitative genetics, we often see the following equation:

\[ P = G + E \]

- We can divide this into total *phenotypic variance*, *genetic variance*, and *environmental variance*:

\[ V_P = V_G + V_E \]

- The total genetic variance divides into additive genetic variance and everything else:

\[ V_P = V_A + V_D + V_I + V_E \]

- This leads to definitions of *narrow sense heritability* and *broad sense heritability*

\[ h^2 = \frac{V_A}{V_P} \quad H^2 = \frac{V_G}{V_P} \]
Heritability concepts V

- Another classic parameterization of genetic effects is the following

\[ G_{A_1A_1} = 0, \quad G_{A_1A_2} = a + d, \quad G_{A_2A_2} = 2a \]

\[ V_A = 2p(1-p)\left(a\left(1 + d(p_1 - p_2)\right)\right)^2 \]

- We can convert these to our regression parameters by solving the following equations and making appropriate substitutions:

\[ 0 = \beta_\mu - \beta_a - \beta_d, \quad a + d = \beta_\mu + \beta_d, \quad 2a = \beta_\mu + \beta_a - \beta_d \]

- Note one last important relationship!:

\[ \beta_\alpha = \beta_a \left(1 + \frac{\beta_d}{2} (p_1 - p_2)\right) \]
Heritability concepts VI

- Change over time depends on the additive genetic variance and the selection gradient:
  \[ \Delta \bar{Y} = h^2 s \]

- Genetic drift depends on the heritability and the effective population size:
  \[ V_{P,t+1} = \frac{h_t^2 V_{P,t}}{N_e} \]

- No heritability means there is no evolution!
Do we still use heritability in quantitative genomics?

- Yes! It’s an important concept for thinking about evolution, the structure of variation in populations, etc.

- It is often important for determining our chances of using a GWAS to map the locations of causal polymorphisms (why is this?)

- We often use marginal heritabilities, i.e. the heritability due to a single marker to provide a quantification of effects (note that we use different concepts such as relative risks and related concepts when dealing with case / control data):

\[ h_m^2 = \frac{2p_i(1 - p_i)\beta_{\alpha,i}^2}{V_P} \]

- In short, heritability is an important concept, but now you have the tools to understand heritability in terms of regressions (!!!) and this will provide a framework for understanding related concepts
Conceptual Overview

- Genetic System
- Sample or experimental pop
- Does A1 → A2 affect Y?
- Measured individuals (genotype, phenotype)
- Reject / DNR
- Regression model
- Model params F-test
- Pr(Y|X)
Conceptual Overview

System

Question

Experiment

Sample

Inference

Distribution

Estimator

Estimator Hypothesis Test

Probability Model

Hypothesis Test
Pep Talk: keep learning (!!)

- How to learn stats, math, coding, etc. (my suggestion):
  - Figure out what you are passionate about (it will involve quantitative aspects!) and build your understanding by hooking it into your passion
  - If you spend more than a few hours / day trying to understand something and can’t it means you are missing a critical component that you have not learned = put it down and come back to it at a later date (you’ll be surprised how you’ll learn something later that suddenly makes it clear…)
  - Don’t memorize theorems, constantly study, etc. = know what you do understand, keep adding to this, and learn it over time by hooking it into your passion
- Don’t be intimidated by others or yourself
  - Anyone trying to make you feel bad because they “know math” and you don’t is confused (knowing math someone else has developed does not mean you’re smart…)
  - I’m too old to learn this, I don’t understand what people are saying / the material in the class so I’m not smart enough to learn it, etc. = NOT TRUE
  - You can stop learning this for extended periods and lose faith in yourself for years - you can always come back and keep learning and you WILL LEARN IT (trust me)
That’s it for today

• That’s it!