Quantitative Genomics and Genetics

BIOCB 4830/6830; PBSB.5201.03

Lecture 28: (Classic) Medical, Agricultural, and Evolutionary Quantitative Genetics

Jason Mezey

May 7, 2024 (T) 8:40-9:55

Announcements

- Last work for the class: Project and Final
  - For project (due TODAY by 11:59PM, May 7!)
  - Final will be same format as midterm (available May 11 and due by 11:59PM, May 18!) and **you will do a GWAS analysis with a linear regression with and without covariates AND a logistic regression with and without covariates (!!)**

Quantitative Genomics and Genetics - Spring 2024
BIOCB 4830/6830; PBSB 5201.01

Final exam available: Sat., May 11

Final exam due: 11:59PM, Sat., May 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 (even ChatGPT or similar!), your own notes and your previously constructed code, etc. **HOWEVER <u>YOU ARE NOT ALLOWED</u> TO COMMUNICATE OR IN ANY <u>WAY ASK ANYONE FOR ASSISTANCE WITH</u> THIS EXAM IN ANY FORM e.g., DO NOT POST PUBLIC MESSAGES ON ED DISCUSSION!** (the only exceptions are Beulah, Sam, and Dr. Mezey, e.g., you MAY send us a private message on Canvas). 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, where you will submit your .Rmd script and the results of running your code in an associated .pdf file (plus an additional .pdf files if you have separate files for your written answers and code output). 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!!).

4. The exam must be uploaded on Canvas before 11:59PM (!!) (ET) Sat, May 18. 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 28: Classic Quantitative Genetics

- Today, we will (briefly) introduce the three fields that are now completely integrated within modern field of Quantitative Genomics / Genetics (Medical Genetics, Agricultural Genetics, and Evolutionary Genetics) and (some) of the methods used by these fields before the introduction of genome-wide genetic data (=GWAS)!

The impact of Genomic Data on genetic analysis

- Before the "Genomic Era" genetic analysis was part of three different fields that used different analysis techniques: **Medical Genetics**, **Agricultural Genetics**, and **Evolutionary Genetics**

- The reason was they were analyzing different systems / interested in different questions AND they did not have the data available to do what they really wanted to do: *identify which differences in a genome (genotypes) were responsible for differences in phenotypes of interest (!!)*

- Once genomic data (i.e., data on the entire genome) became available the starting analysis of all of these fields became the same (i.e., analyzing which differences impacted phenotypes) *and they started using the same set of methods (!!)* = effectively unifying these fields into modern "Quantitative Genetics / Genomics"

- This is the reason the Quantitative Genetics literature before the Genomic Era is so difficult to follow / seems so diffuse… but after this class you will understand how to go back and figure out this literature (!!)

A few definitions I

- **Association analysis** - any analysis involving a statistical assessment of a relation between genotype and phenotype, e.g. a hypothesis test involving a multiple regression model
- **Mapping analysis** - an association analysis
- **Linkage disequilibrium (LD) mapping** - an association analysis
- **Segregating** - any locus where there is more than one allele in the population
- **Genetic marker** - any segregating polymorphism we have measured in a GWAS, i.e. SNPs genotyped in a GWAS
- **Tag SNP** - a SNP correlated with a causal polymorphism
- **Locus** or **Genetic Locus** - a position in the genome (which may refer to a single polymorphism or an entire genomic segment, e.g. that contains the coding region of a gene

A few definitions II

- **Mendelian trait** - any phenotype largely affected by one or at most two loci where environment does not have a large effect on the phenotype
- **Complex trait** - any phenotype affected by more than one or two loci and/or where environmental effects account for most of the variation we observe in a population
- **Quantitative trait** - a complex trait

A few definitions III

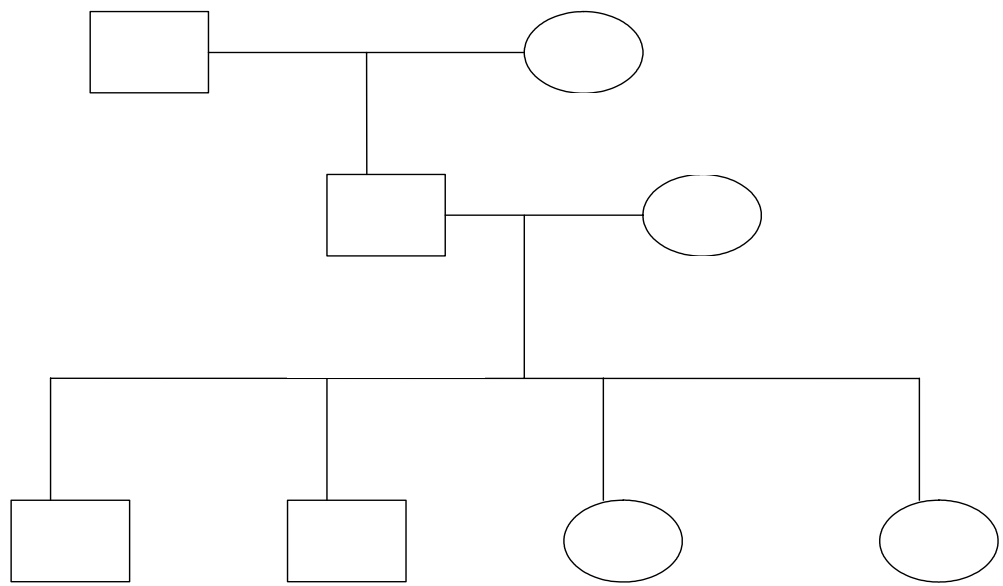
- **Quantitative Trait Locus (QTL)** - a causal polymorphism (or the locus containing the polymorphism) OR a large section of the the genome containing a causal (or several!) polymorphisms
- **expression Quantitative Trait Locus (eQTL)** - a QTL for a gene expression phenotype, i.e. a quantitative measurement 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)

(Medical Genetics) Association analysis when samples are from a pedigree

- The "ideal" GWAS experiment is a sampling experiment where we assume that the individuals meet our i.i.d. assumption
- There are many ways (!!) that a sampling experiment does not conform to this assumption, where we need to take these possibilities into account (what is model we have applied in this type of case?)
- Relatedness among the individuals in our sample is one such case
- This is sometimes a nuisance that we want to account for in our GWAS analysis (what is an example of a technique used if this is the case?)
- It is also possible that we have sampled related individuals ON PURPOSE because we can leverage this information (if we know how the individuals are related...) using specialized analysis techniques (which have a GWAS analysis at their core!)
- Analysis of pedigrees is one such example, where inbred lines (a special class of pedigrees!) is another

What is a pedigree?

- **pedigree -** a sample of individuals for which we have information on individual relationships
- Note that this can cover a large number of designs (!!), i.e. family relationships, controlled breeding designs, more distant relationships, etc.
- Standard representation of a family pedigree (females are circles, males are squares):



Pedigrees in genetics I

- Use of pedigrees has a long history in genetics, where the use of family pedigrees stretch back ~100 years, i.e. before genetic markers (!!)
- The observation that lead people to analyze pedigrees was that Mendelian diseases (= phenotype determined by a single locus where genotype is highly predictive of phenotype) tend to run in families
- The genetics of such diseases could therefore be studies by analyzing a family pedigree
- Given the disease focus, it is perhaps not surprisingly that family pedigree analysis was the main tool of medical genetics

Pedigrees in genetics II

- When the first genetic markers appeared, it was natural to use these to identify positions in the genome that may have the causal polymorphisms responsible for the Mendelian disease
- In fact, analysis of pedigrees in combination with just a few markers was the first step in identifying the causal polymorphisms for many Mendelian diseases, i.e. they could identify the general position in a chromosome, which could be investigator further with additional markers, tec.
- In the late 70's - 90's a large number of Mendelian causal disease polymorphisms were found using such techniques
- Pedigree analysis therefore dominates the medical genetics literature (where now this field is wrapped into the more diffusely field of quantitative genomics!)

Types of pedigree analysis

- **segregation analysis** - inference concerning whether a phenotype (disease) is consistent with a Mendelian disease given a pedigree (no genetic data!)
- **identity by descent (ibd)** - inference concerning whether two individuals (or more) individuals share alleles because they inherited them from a common ancestor (note: such analyses can be performed without markers but more recently, markers have allowed finer ibd inference and ibd inference without a pedigree!)
- **linkage analysis** - use of a genetic markers on a pedigree to map the position of causal polymorphisms affecting a phenotype (which may be Mendelian or complex)
- **family based testing** - the use of genetic markers and many small pedigrees to map the position of causal polymorphisms (again Mendelian or complex)
- Note that there are others (!!) and that we will provide simple example to the illustrate linkage analysiis

Importance of pedigree analysis now

- The reason that we do not focus on pedigree analysis in this class is the having high-coverage marker data makes many of the pedigree analyses unnecessary
- As an example, pedigree (linkage) analysis was useful when we only had a few markers because we could use the pedigree to infer the states of unseen markers
- Once we can measure all the markers there is no need to use a pedigree
- In fact, we can easily map the positions of Mendelian disease causal polymorphisms without a pedigree (and we now do this all the time)
- What's worse, using pedigree (linkage) analysis to map causal polymorphisms to complex phenotypes are turning out to have produced more (=not useful) inferences (!!)
- However, understanding the basic intuition of these methods is critical for understanding the literature in quantitative genetics, for cases where the sample has to be from a pedigree (e.g., for a rare genetic disease in just a few families) and for derived pedigree methods that are used in GWAS

Connection between linkage / association analysis I

- Both linkage analysis and association analysis have the same goal: identify positions in the genome where there are causal polymorphisms using genetic markers
- Recall that we are modeling the following in association analysis:

$$Pr(Y|X)$$

- We are not concerned that the marker we are testing is not the causal marker, but we would prefer to test the causal marker (if we could!)
- Note that if we could model the relationship of the unmeasured causal polymorphism X_{cp} and observed genetic marker X, we could use this information:

$$Pr(Y|X_{cp})Pr(X_{cp}|X)$$

- This is what we do in linkage analysis (!!)

Connection between linkage / association analysis II

- Note that the first of these two terms is called the *penetrance* model (and there are many ways to model penetrance!) and the second term is modeled based on the structure of an observed pedigree, which allows us to infer the conditional relationship of the causal polymorphism and observed genetic marker by inferring a recombination probability parameter *r* (confusingly, this is often symbolized as in the literature!): θ

$$Pr(Y|X_{cp})Pr(X_{cp}|X, r_{(X_{cp},X)})$$

- We can therefore use the same statistical (inference) tools we have used before but our models will be a little more complex and we will be inferring not only parameters that relate the genotype and phenotype (e.g. regression β 's) but also the parameter *r* (!!)

- If we are dealing with a Mendelian trait (which is the case for many linkage analyses), the causal polymorphism perfectly describes the phenotype so we do not need to be concerned with the penetrance model:

$$Pr(X_{cp}|X, r_{(X_{cp},X)})$$

Connection between linkage / association analysis III

- In the literature, we often symbolize the combination of Xcp and X as a single g (for the genotype involving both of these polymorphisms) so we may re-write this equation as the probability of a vector of a sample of n of these genotypes:

$$Pr(\mathbf{X}_{cp}|\mathbf{X}, r) = Pr(\mathbf{g}|r)$$

- To convert this probability model into a more standard pedigree notation, note that we can write out the genotypes of the n individuals in the sample

$$Pr(g_1, ..., g_n|r)$$

- Using the pedigree information, we can write the following conditional relationships relating parents (father = g_f, mother = g_m) to their offspring (where individuals without parents in the pedigree are called *founders*):

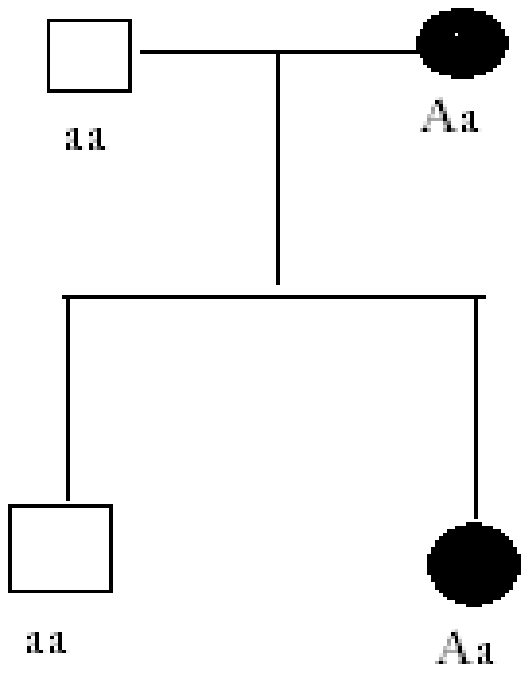
$$\prod_{i}^{f} Pr(g_i) \prod_{j=f+1}^{n} Pr(g_j|, g_{j,f}, g_{j,m}, r)$$

- Finally, for inference, we need to consider all possible genotype configurations that could occur for these n individuals (=classic pedigree equation):

$$\sum_{\Theta_g} \prod_{i}^{f} Pr(g_i) \prod_{j=f+1}^{n} Pr(g_j|, g_{j,f}, g_{j,m}, r)$$

Simple linkage analysis example I

- Consider the following pedigree where we have observed a marker allele with two states (A and a) and the phenotype healthy (clear) and disease (dark) where we know this is a Mendelian disease where the disease causing allele D is dominant to the healthy allele (i.e. individuals who are DD or Dd have the disease, individuals who are dd are healthy) and is very rare (such that we only expect one of these alleles in this family):



Simple linkage analysis example II

- For this example, the probability model is as follows:

$$\sum_{\Theta_g} \prod_{i}^{f} Pr(g_i) \prod_{j=f+1}^{n} Pr(g_j|, g_{j,f}, g_{j,m}, r) = \sum_{\Theta_g} Pr(g_f) Pr(g_m) Pr(g_1|g_f, g_m) Pr(g_2|g_f, g_m)$$

- Given what we know about the system, there are two possible genotype configurations (why?):

$$\Theta_g = \{\{ad/ad, AD/ad, ad/ad, AD/ad\}, \{ad/ad, Ad/aD, ad/ad, AD/ad\}\}$$

- If we assign p1(A) = frequency of A, p2(D) = frequency of D, and we assume Hardy-Weinberg frequencies for the founders (which we often do in pedigree analyses!) we get:

$$Pr(g_f) Pr(g_m) = ((1-p_1)^2 * (1-p_2)^2)(2p_1(1-p_1) * 2p_2(1-p_2)) = 4p_1p_2(1-p_1)^3(1-p_2)^3$$

- Note there are two possible configurations for the genotypes of the offspring:

$$Pr(g_1|g_f, g_m) Pr(g_2|g_f, g_m) = Pr(ad/ad|ad/ad, AD/ad) Pr(AD/ad|ad/ad, AD/ad) = \frac{1-r}{2}\frac{1-r}{2}$$

$$Pr(g_1|g_f, g_m) Pr(g_2|g_f, g_m) = Pr(ad/ad|ad/ad, Ad/aD) Pr(AD/ad|ad/ad, Ad/aD) = \frac{r}{2}\frac{r}{2}$$

- Putting this together, we get the following probability model for this case:
$\sum_{\Theta_g} Pr(g_f) Pr(g_m) Pr(g_1|g_f, g_m) Pr(g_2|g_f, g_m) = p_1p_2(1-p_1)^3(1-p_2)^3[(1-r)^2 + r^2]$

Simple linkage analysis example III

- Note that this probability model defines a likelihood (!!) such that we can perform a likelihood ratio test for whether the marker is in LD with the disease (causal) polymorphism (we can also do this in a Bayesian framework!)
- The actual hypothesis we would test in this simple Mendelian case is that H0: r = 0.5 with HA: r any value between 0 and 0.5 (why is this?)
- For complex phenotypes, we could also have a regression (glm!) model as part of our likelihood and therefore likelihood ratio test
- Note that calculating likelihood (or posteriors!) for complex pedigrees gets very complicated (think of all the genotype configurations!) requiring algorithms, many of which are classics (and implemented in pedigree analysis software), i.e. peeling algorithm, etc.
- Also note, that many of these programs consider models with more than one marker at a time, i.e. multi-point analysis

Linkage analysis wrap-up

- Again, note that in general, linkage analysis provides useful information when you have a Mendelian phenotype and low marker coverage
- If you have a more complex phenotype or higher marker coverage, it is better just to test each marker one at a time, since the additional model complexities in linkage analysis tend to reduce the efficacy of the inference
- A downside of using pedigrees designs for mapping with high marker coverage is they have high LD (why?) so resolution is low
- An upside is the individuals in the sample can be enriched for a disease (particularly important if the disease is rare) and by considering individuals in a pedigree, this provides some control of genetic background (e.g. epistasis) and other issues!
- This latter control is why family-based tests are also still used

Family based tests I

- There are a large number of family based testing methods for mapping causal polymorphisms
- While each of these work in slightly different ways, each calculates a statistic based on the association of a genetic marker with a disease phenotype for sets of small families (=the family, not the individual is the unit), i.e. trios, nuclear families, etc.
- These statistics are then used to assess whether the marker is being transmitted in each family with the disease in a hypothesis testing framework (null hypothesis = no co-transmission), where rejection of the null indicates that the marker is in LD with a causal polymorphism
- An advantage of using family based tests is treating the family as a unit controls for covariates (e.g. population structure) although the downside is smaller sample size n because individuals are grouped into families (why is this a downside?)
- If you have a design which allows family based testing, a good rule is to apply both family based tests and standard association tests (that we have learned in this class!)

Family based tests II

- As an example, there are many family based tests in the Transmission-Disequilibrium Testing (TDT) class

- These generally use trios (parents and an offspring) counting the cases where which chromosome is transmitted from a parent is clear and whether the case was affected or unaffected:

Parent 1	Parent 2	Affected	Unaffected	b	c
Aa	Aa	AA	Aa	1	0
Aa	Aa	Aa	aa	1	1
Aa	aa	Aa	aa	1	0
AA	AA	AA	AA	0	0
Aa	Aa	AA	Aa	2	0
Aa	aa	Aa	aa	1	0
aa	Aa	aa	aa	0	1
Aa	Aa	AA	aa	2	0
Aa	AA	AA	Aa	1	0
AA	aa	Aa	Aa	0	0
			Sum:	9	2

- The test statistic is the a z-test (look it up on wikipedia!)

$$Z_{TDT} = \frac{b - c}{\sqrt{b + c}}$$

(Agricultural Genetics) 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

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 (!!) (use a mixed model to account for inbred line structure…)

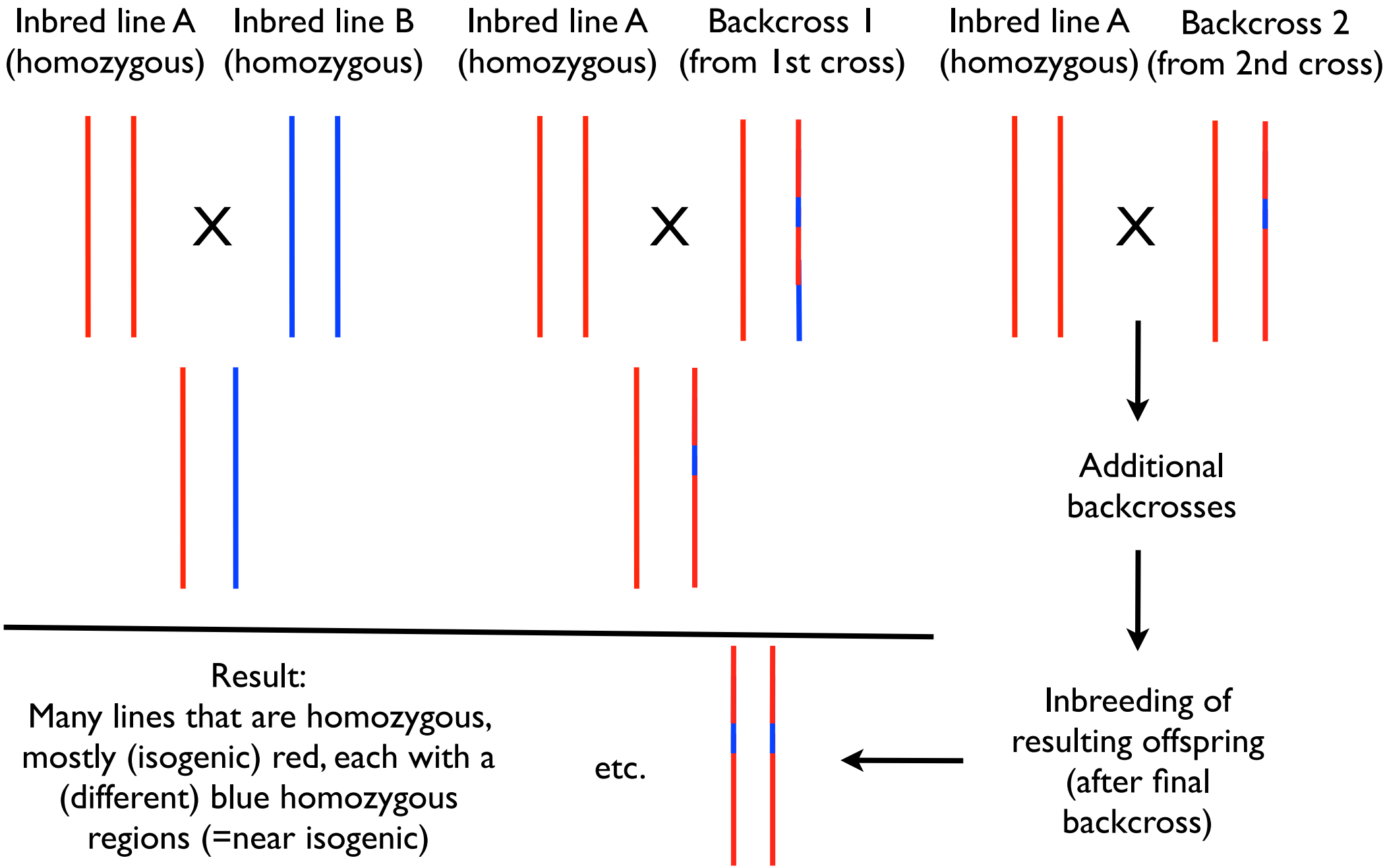
Types of inbred line designs (important in genetic analysis)

- 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

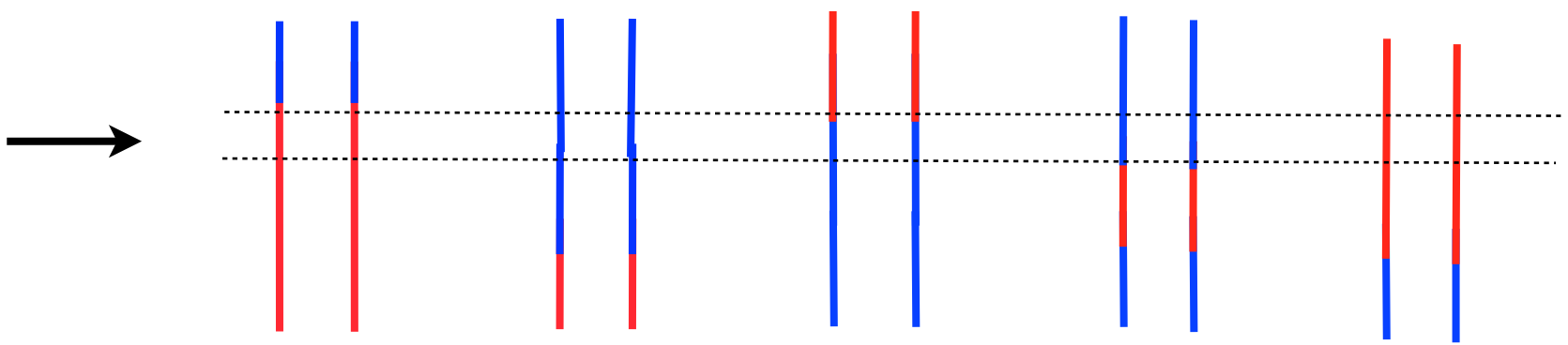
- The reason that inbred line designs are useful is we can infer the unobserved markers (with low error!) even with very few markers
- The reason is inbred lines designs result in homozygosity of the resulting lines (although they may be homozygous for different genotype!)
- Therefore, inbreeding, in 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



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!):

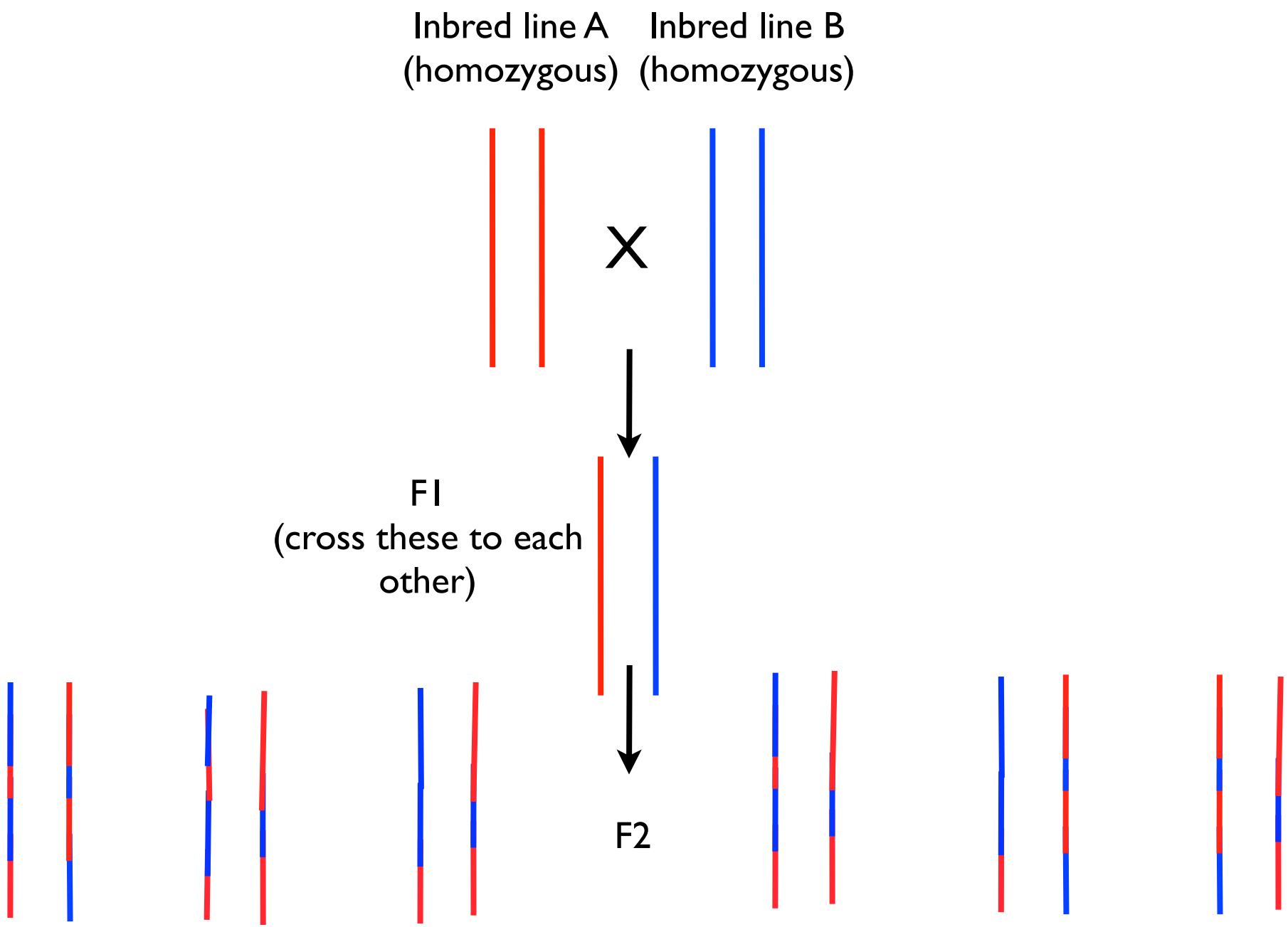
$$Pr(Y|X_{cp=Q})Pr(X_{cp}|X, r_{(X_{cp=Q},X)}) = \sum_{\Theta_g} \prod_{i}^{f} Pr(y|g_i)Pr(g_i)Pr(g_i) \prod_{j=f+1}^{n} Pr(y_j|g_j)Pr(g_j|, g_{j,f}, g_{j,m}, r)$$

- 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:

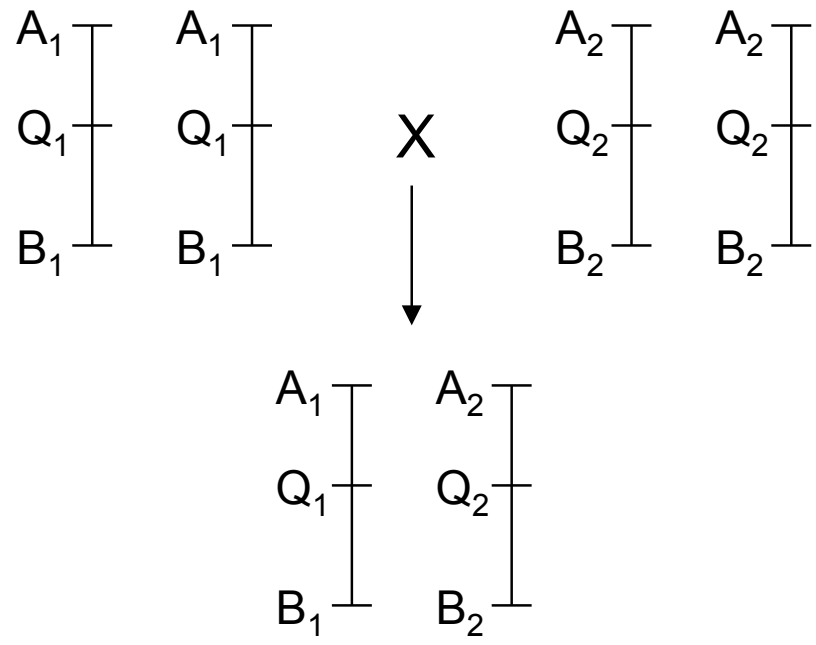
$$Pr(Y|X_{cp=Q})Pr(X_{cp}|X, r_{(X_{cp=Q},X)}) = \prod_{i}^{n} \sum_{\Theta_g} Pr(y_i|g_{i,Q})Pr(g_{i,Q}|g_{i,A}, g_{i,B}, r)$$

- 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

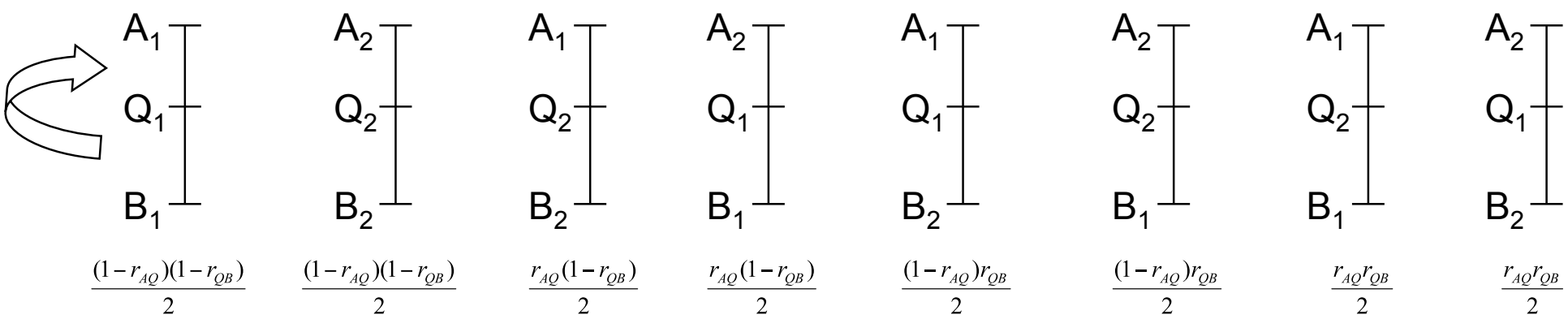
Example 2: interval mapping (F2)



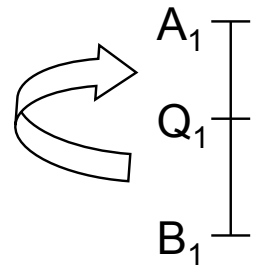
Example 2: interval mapping (F2)



F1 Gametes:

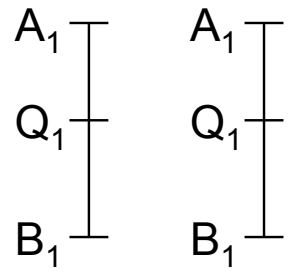


Example 2: interval mapping (F2)



$$\frac{(1-r_{AQ})(1-r_{QB})}{2}$$

F2:



$$\left(\frac{(1-r_{AQ})(1-r_{QB})}{2}\right)^2$$

$$\Pr(Q_1Q_1 \mid A_1A_1B_1B_1) = \frac{\Pr(A_1A_1Q_1Q_1B_1B_1)}{\Pr(A_1A_1B_1B_1)} = \frac{\left(\frac{(1-r_{AQ})(1-r_{QB})}{2}\right)^2}{\left(\frac{1-r_{AB}}{2}\right)^2} = \frac{(1-r_{AQ})^2(1-r_{QB})^2}{(1-r_{AB})^2}$$

Example 2: interval mapping (F2)

- We can therefore substitute these conditional probabilities into our main equation and calculate the likelihood over possible values of r

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

- 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!)

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 (!!)

(Evolutionary Genetics) heritability and additive genetic variance

- 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:

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

- 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):

$$V_P = \frac{1}{n}\sum_{i}^{n}(Y_i - \bar{Y})^2$$

- 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:

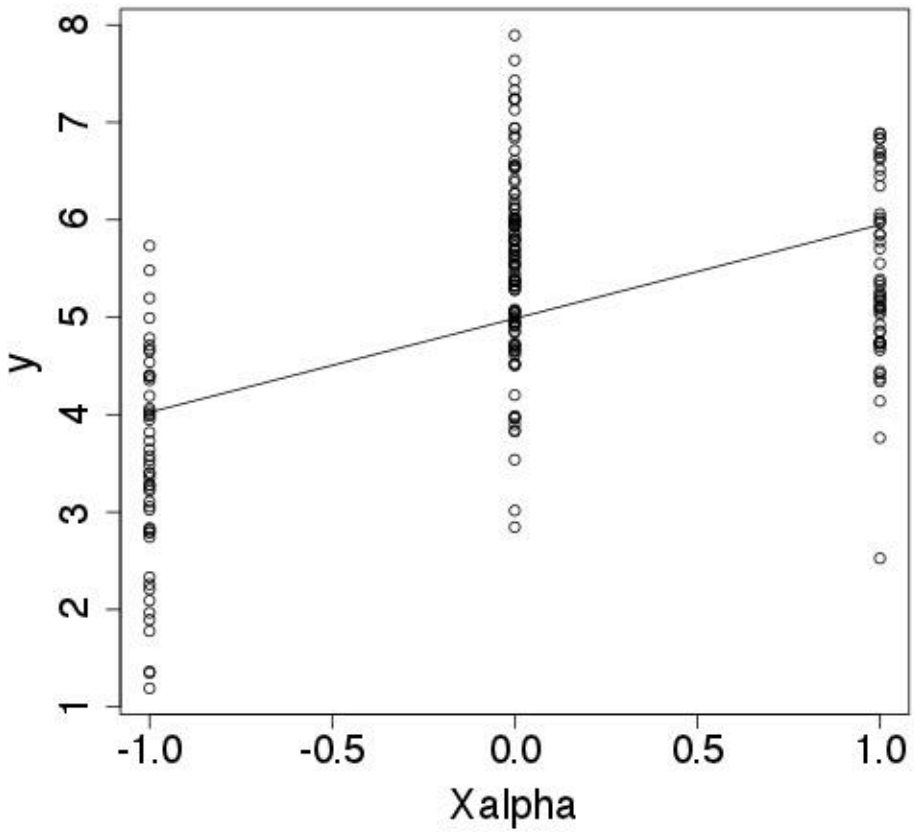
$$V_A = 2p(1-p)\beta_\alpha^2$$

Additive genetic variance II

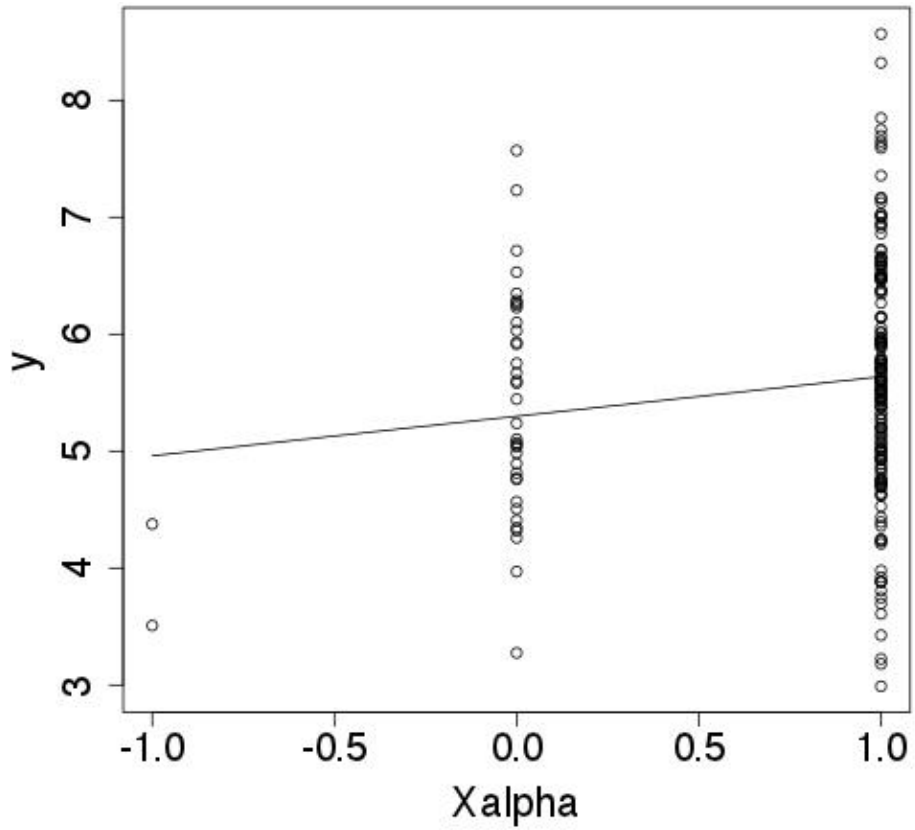
- There is a consequence of whether we fit two or one "slope" parameters in our regression model
  - If we consider two slope parameters β_a, β_d (as we have done all semester!) the true values of the parameters are the same regardless of the allele frequency (MAF) of the causal polymorphism
  - If we consider one regression parameter β_α 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: $Y = \beta_\mu + X_a\beta_\alpha + \epsilon$



MAF=0.5, larger β_α

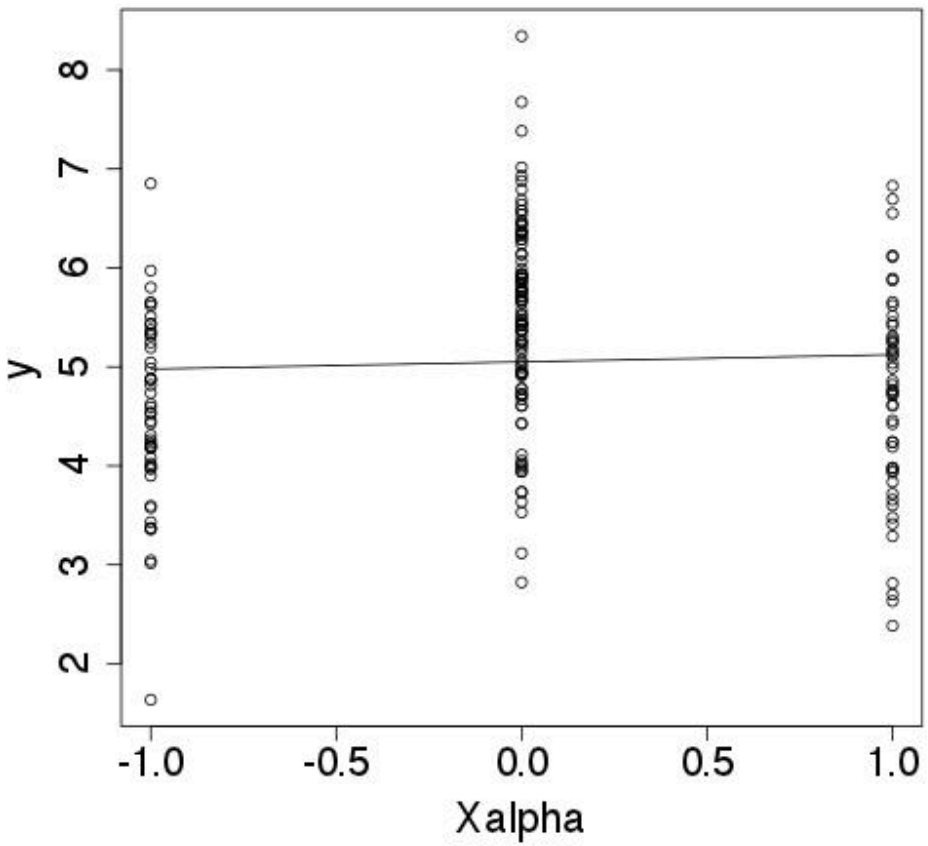


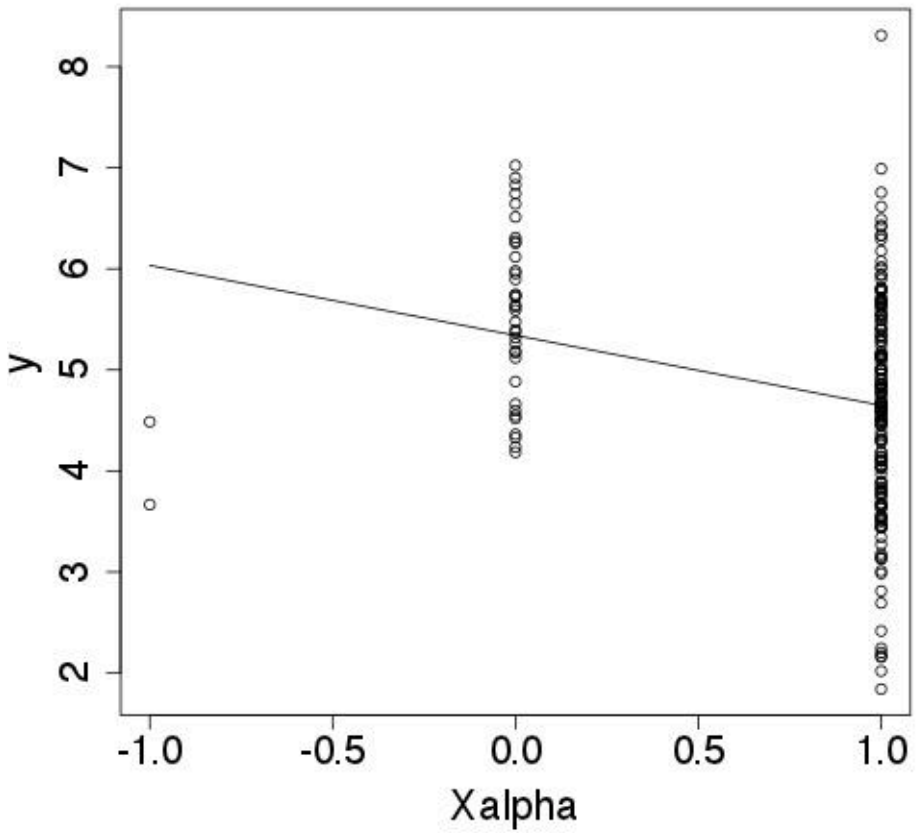
MAF=0.1, smaller β_α

- Remember (!!) this is not the case if we fit two parameters: β_a, β_d

Example how the parameter changes with MAF I

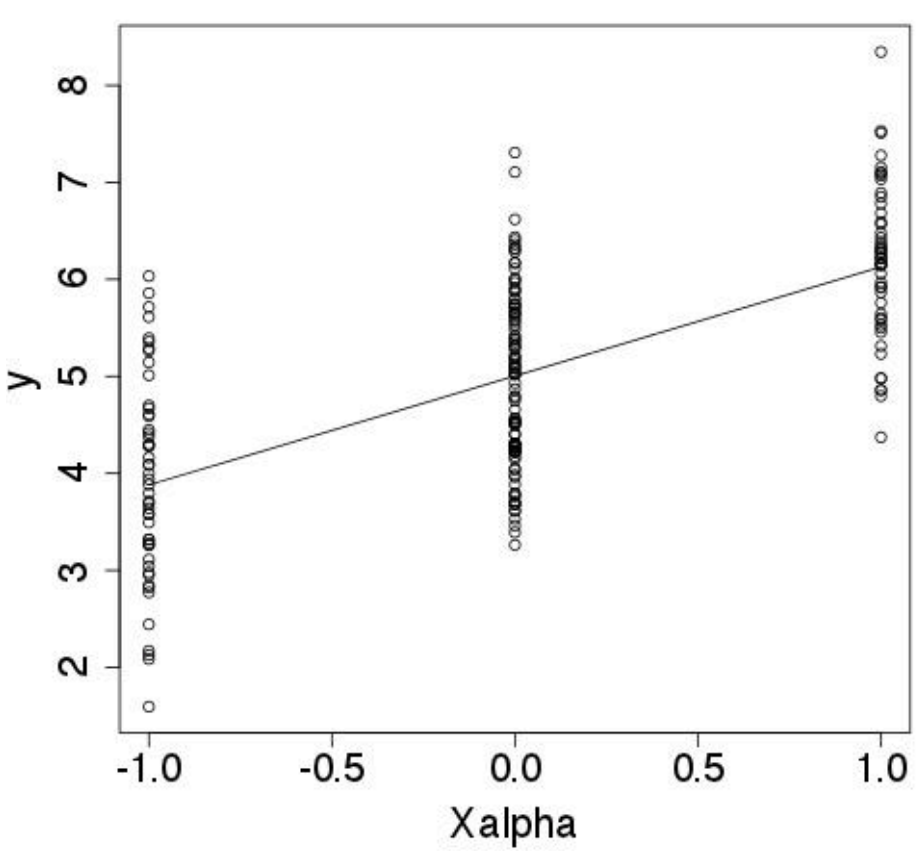
- In a case of over-dominance (or under-dominance) with the right allele frequency, the true value of the parameter can be zero (!!):

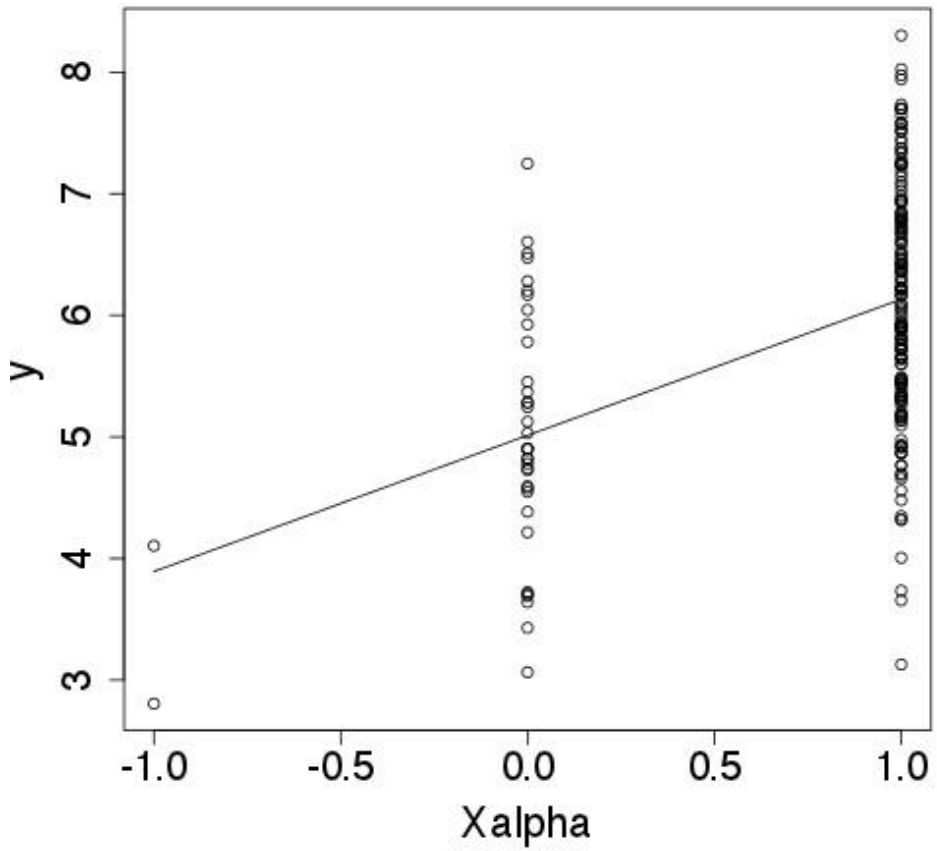




Example how the parameter changes with MAF III

- In a purely additive case (no dominance) the parameter β_α does not change, regardless of MAF:





- This makes sense since we only need the parameters β_μ, β_α 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 β_α 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:

$$h^2 = \frac{V_A}{V_P} = \frac{2p(1-p)\beta_\alpha^2}{V_P}$$

- 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 I

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

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

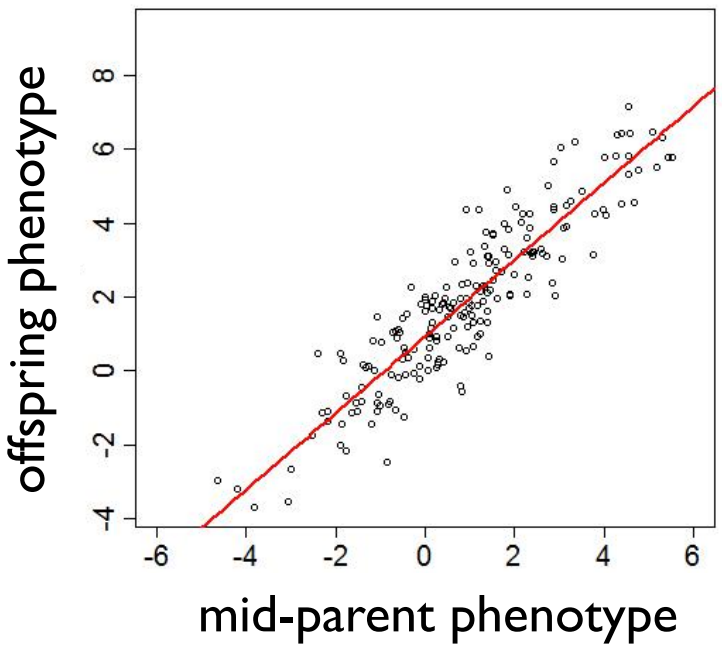
- The equations get more complex for LD and epistasis (and for more alleles, etc.

- 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*

$$h^2 = \frac{V_A}{V_P} \qquad H^2 = \frac{V_G}{V_P}$$

Heritability concepts V

- Another classic parameterization of genetic effects is the following

$$G_{A_1A_1} = 0,\ G_{A_1A_2} = a + d,\ G_{A_2A_2} = 2a$$

$$V_A = 2p(1-p)\Big(a\Big(1 + d(p_1 - p_2)\Big)\Big)^2$$

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

$$0 = \beta_\mu - \beta_a - \beta_d,\ a + d = \beta_\mu + \beta_d,\ 2a = \beta_\mu + \beta_a - \beta_d$$

- 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:

$$V_{\bar{P},t+1} = \frac{h_t^2 V_{P,t}}{N_e}$$

- No heritability means there is no evolution!

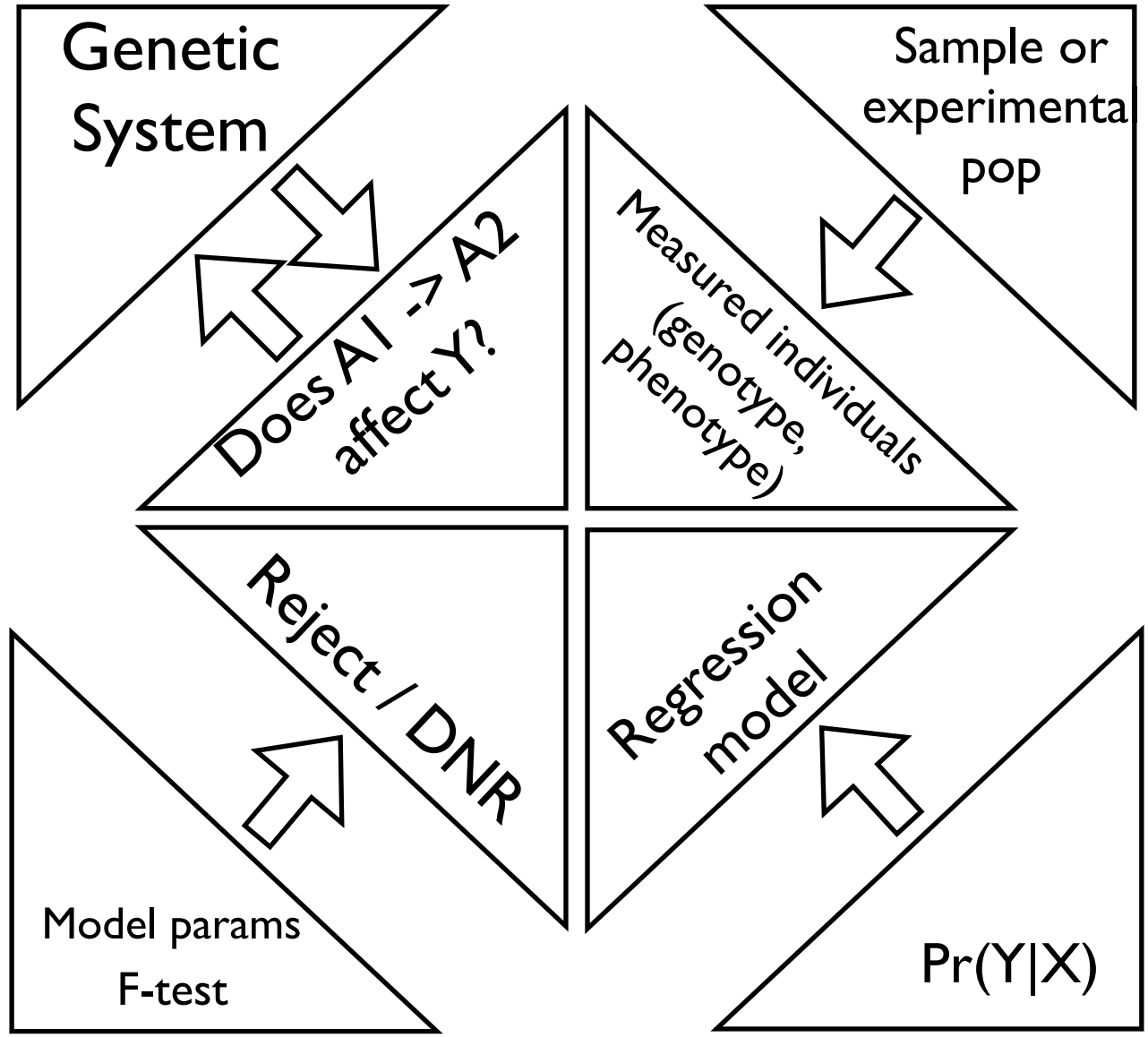
Do we still use heritability in quantitative genomics?

- 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 polymorphims (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):

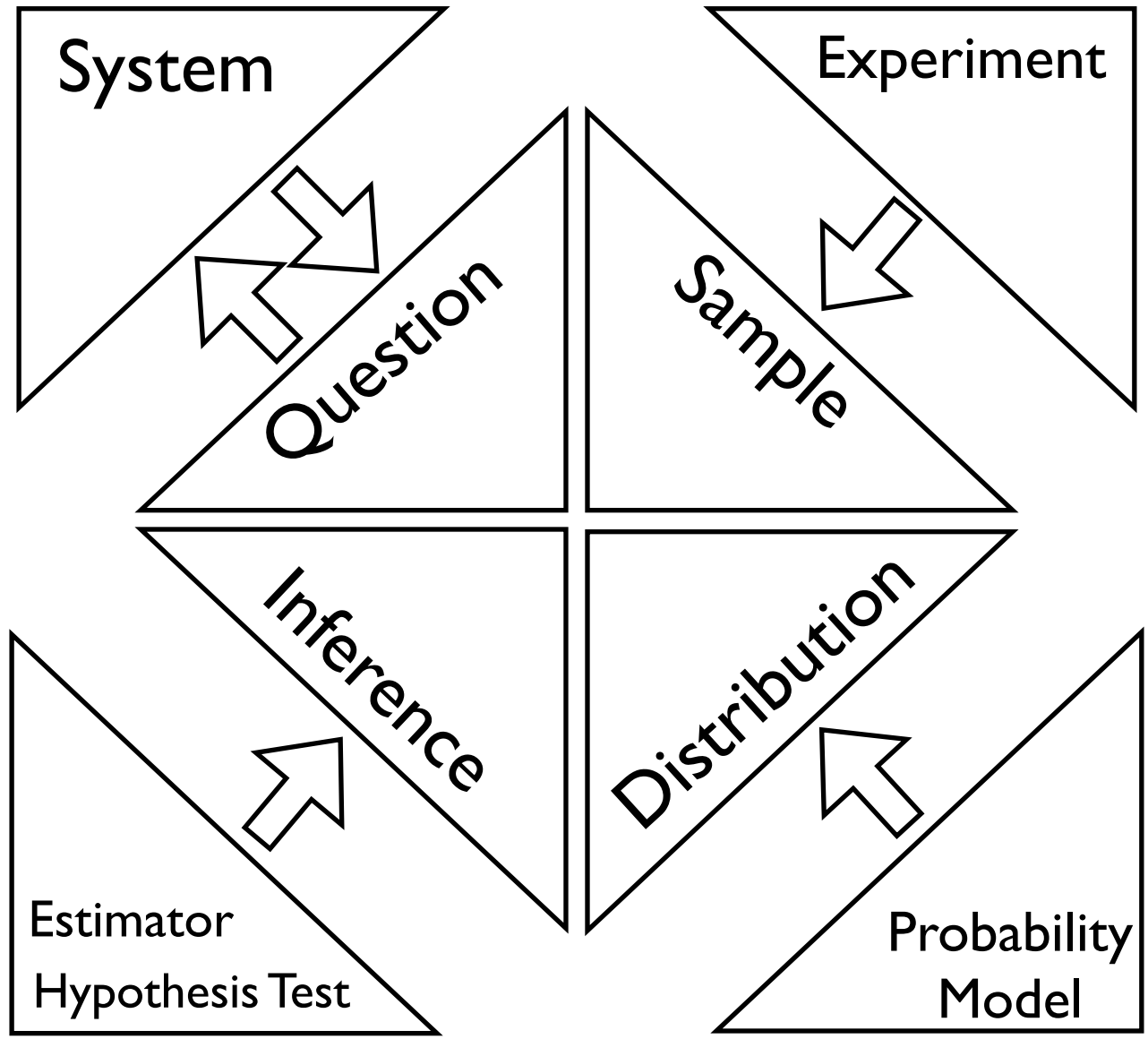
$$h^2_m = \frac{2p_i(1-p_i)\beta^2_{\alpha,i}}{V_P}$$

- 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



Conceptual Overview



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 (prioritize intuition!!), 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! Best wishes to you all!