Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

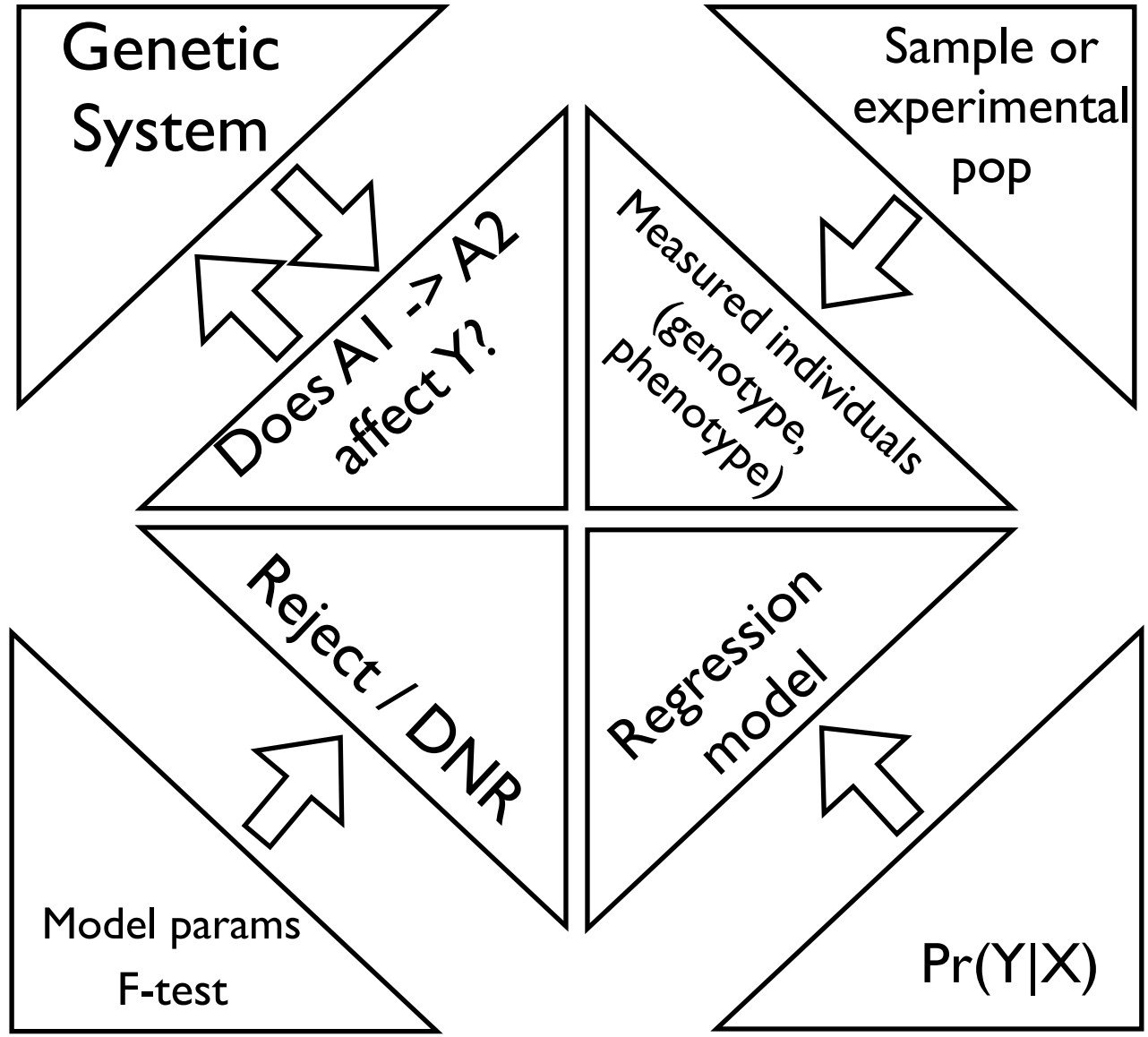
Optional Lecture 2: Multiple genotypes and phenotypes

Jason Mezey
jgm45@cornell.edu
May 3, 2021 (M) 2:30-3:30PM

Summary of Optional Lecture 2

- Today we will discuss how to incorporate multiple genotypes into the linear regression model and testing for epistasis (i.e., genetic interactions = interactions between genotypes)
- We'll also discuss analysis of multiple phenotypes and a particularly important example: expression Quantitative Trait Loci (=eQTL)

Conceptual Overview



Introduction to epistasis I

- So far, we have applied a GWAS analysis by considering statistical models between one genetic marker and the phenotype
- This is the standard approach applied in all GWAS analyses and the one that you should apply as a first step when analyzing GWAS data (always!)
- However, we could start considering more than one marker in each of the statistical models we consider
- One reason we might want to do this is to test for statistical interactions among genetic markers (or more specifically, between the causal polymorphisms that they are tagging)

Introduction to epistasis II

- If we wanted to consider two markers at a time, our current statistical framework extends easily (note that a index AFTER a comma indicates a different marker):

$$Y = \gamma^{-1}(\beta_\mu + X_{a,1}\beta_{a,1} + X_{d,1}\beta_{d,1} + X_{a,2}\beta_{a,2} + X_{d,2}\beta_{d,2}) + \epsilon$$

- However, this equation only has four regression parameters and with two markers, we have more than four classes of genotypes

- To make this explicit, recall that we define the genotypic value of the phenotype as the expected value of the phenotype Y given a genotype:

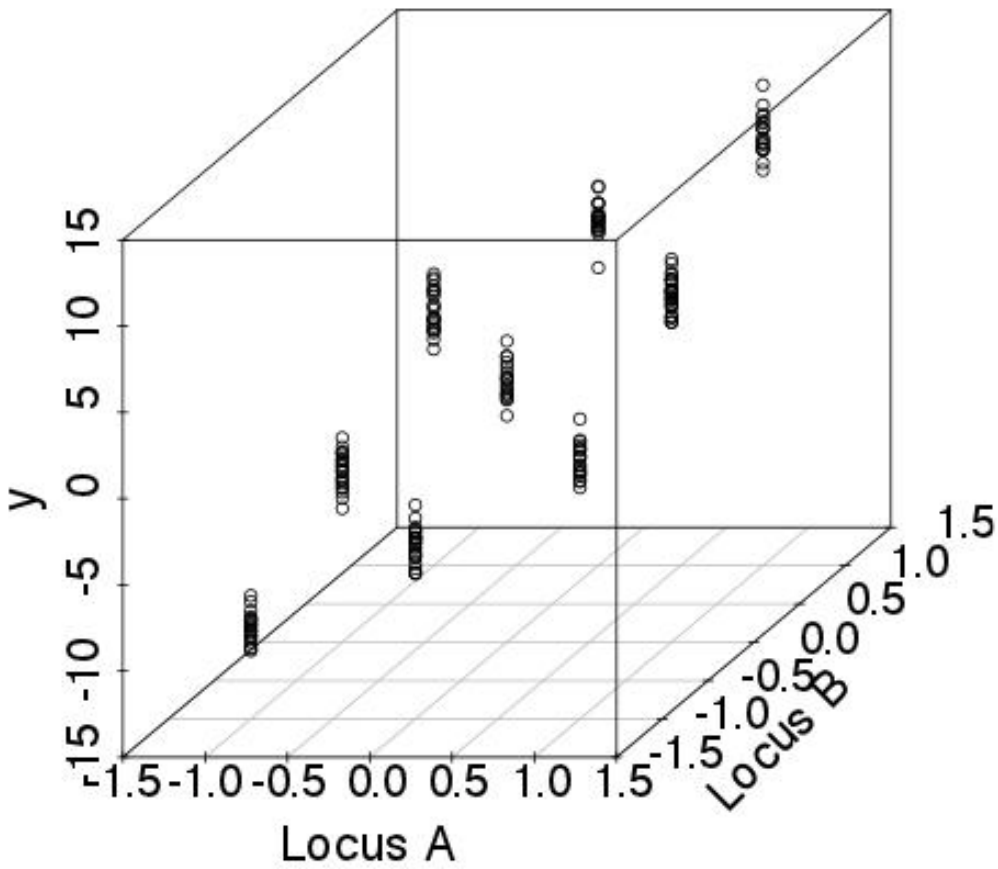
$$G_{A_k A_l B_k B_l} = E(Y|g = A_k A_l B_k B_l)$$

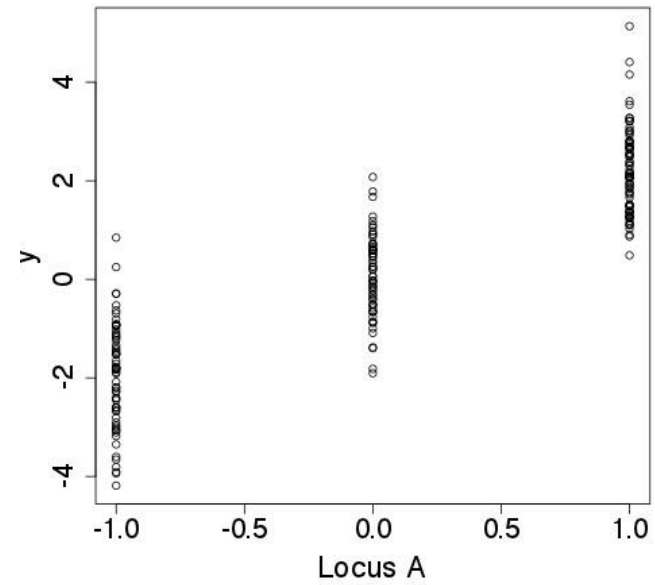
- For the case of two markers, we therefore have nine classes of genotypes and therefore nine possible genotypic values, i.e. we need nine parameters to model this system (why are there nine?):

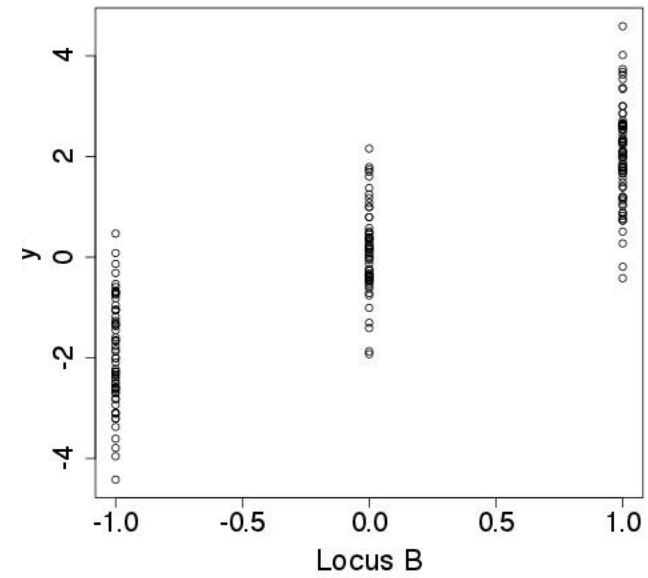
	B_1B_1	B_1B_2	B_2B_2
A_1A_1	$G_{A_1A_1B_1B_1}$	$G_{A_1A_1B_1B_2}$	$G_{A_1A_1B_2B_2}$
A_1A_2	$G_{A_1A_2B_1B_1}$	$G_{A_1A_2B_1B_2}$	$G_{A_1A_2B_2B_2}$
A_2A_2	$G_{A_2A_2B_1B_1}$	$G_{A_2A_2B_1B_2}$	$G_{A_2A_2B_2B_2}$

Introduction to epistasis III

- As an example, for a sample that we can appropriately model with a linear regression model, we can plot the phenotypes associated with each of the nine classes:



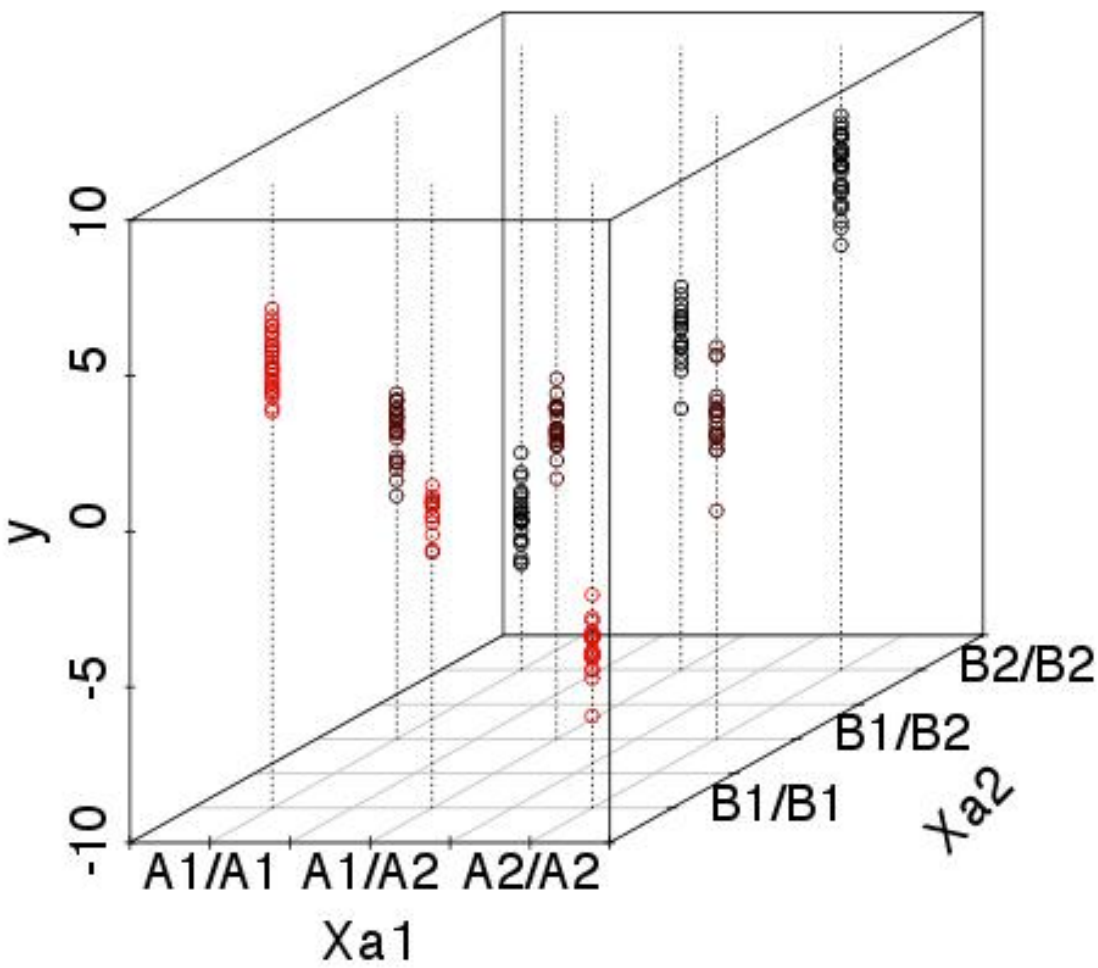


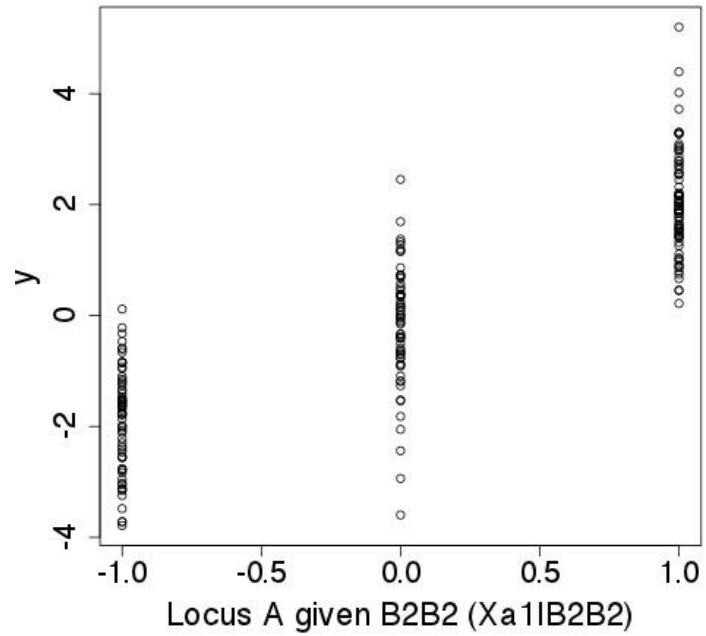


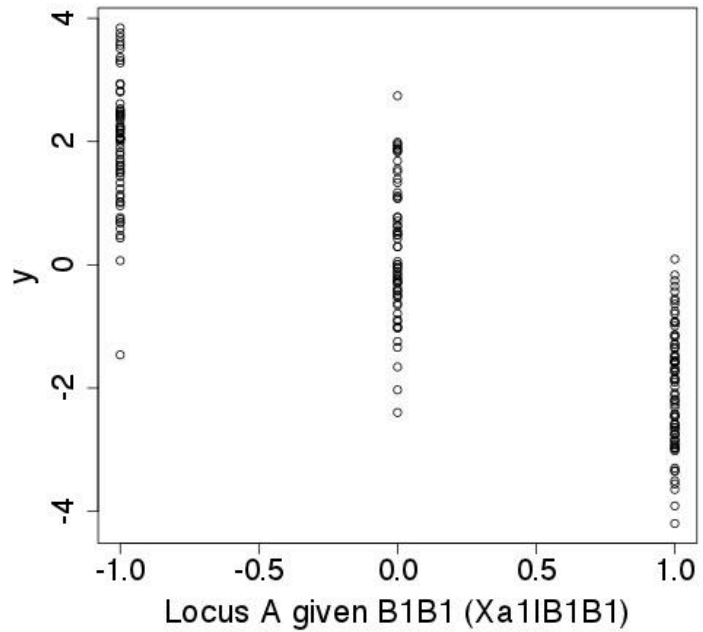
- In this case, both marginal loci are additive

Introduction to epistasis IV

- With nine classes, we also get the possibility of conditional relationships we have not seen before:







- This is an example of *epistasis*

Notes about epistasis 1

- **epistasis** - a case where the effect of an allele substitution at one locus A1 -> A2 alters the effect of a substituting an allele at another locus B1->B2
- This may be equivalently phrased as a change in the expected phenotype (genotypic value) for a genotype at one locus conditional on the state of a locus at another marker
- Note that there is a symmetry in epistasis such that if the effect of at least one allelic substitution (from one genotype to another) for one locus depends on the genotype at the other locus, then at least one allelic substitution of the other locus will be dependent as well
- A consequence of this symmetry is if there is an epistatic relationship between two loci BOTH will be causal polymorphisms for the phenotype (!!!)
- If there is an epistatic effect (=relationship) between loci, we would therefore like to know this information
- Note that we need not consider such relationships for a pair of loci, but such relationships can exist among three (three-way), four (four-way), etc.
- The amount of epistasis among loci for any given phenotype is unknown (but without question it is ubiquitous!!)

Notes about epistasis II

- Note that the definition of epistasis is entirely statistical (!!) and says nothing about mechanism (although people have mis-appropriated the term in this way)
- The term epistasis was coined by Fisher in the 1920's
- Epistasis is sometimes called genotype by genotype, G by G, or G x G
- Geneticists often use the term "modifiers" to describe the dependence of genetic effects at a locus on the state of another locus - this is just epistasis (!!)
- We can also consider the effects of a locus when considering the entire "genetic background" (i.e. all the state in the rest of the genome!) - this is also epistasis (!!)

Modeling epistasis I

- To model epistasis, we are going to use our same linear regression framework (!!)
- The parameterization (using Xa and Xd) that we have considered so far perfectly models any case where there is no epistasis
- We will account for the possibility of epistasis by constructing additional dummy variables and adding additional parameters (so that we have 9 total)

Modeling epistasis II

- Recall the dummy variables we have constructed so far:

$$X_{a,1} = \begin{cases} -1 & \text{for } A_1A_1 \\ 0 & \text{for } A_1A_2 \\ 1 & \text{for } A_2A_2 \end{cases} \qquad X_{d,1} = \begin{cases} -1 & \text{for } A_1A_1 \\ 1 & \text{for } A_1A_2 \\ -1 & \text{for } A_2A_2 \end{cases}$$

$$X_{a,2} = \begin{cases} -1 & \text{for } B_1B_1 \\ 0 & \text{for } B_1B_2 \\ 1 & \text{for } B_2B_2 \end{cases} \qquad X_{d,2} = \begin{cases} -1 & \text{for } B_1B_1 \\ 1 & \text{for } B_1B_2 \\ -1 & \text{for } B_2B_2 \end{cases}$$

- We will use these dummy variables to construct additional dummy variables in our linear regression (and add additional parameters) to account for epistasis A1A1B1B1

$$Y = \gamma^{-1}(\beta_\mu + X_{a,1}\beta_{a,1} + X_{d,1}\beta_{d,1} + X_{a,2}\beta_{a,2} + X_{d,2}\beta_{d,2} + X_{a,1}X_{a,2}\beta_{a,a} + X_{a,1}X_{d,2}\beta_{a,d} + X_{d,1}X_{a,2}\beta_{d,a} + X_{d,1}X_{d,2}\beta_{d,d})$$

Modeling epistasis III

$$Y = \gamma^{-1}(\beta_\mu + X_{a,1}\beta_{a,1} + X_{d,1}\beta_{d,1} + X_{a,2}\beta_{a,2} + X_{d,2}\beta_{d,2} + X_{a,1}X_{a,2}\beta_{a,a} + X_{a,1}X_{d,2}\beta_{a,d} + X_{d,1}X_{a,2}\beta_{d,a} + X_{d,1}X_{d,2}\beta_{d,d})$$

- To provide some intuition concerning what each of these are capturing, consider the values that each of the genotypes would take for dummy variable Xa,1:

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	-1	-1	-1
A_1A_2	0	0	0
A_2A_2	1	1	1

Modeling epistasis IV

$$Y = \gamma^{-1}(\beta_\mu + X_{a,1}\beta_{a,1} + X_{d,1}\beta_{d,1} + X_{a,2}\beta_{a,2} + X_{d,2}\beta_{d,2} + X_{a,1}X_{a,2}\beta_{a,a} + X_{a,1}X_{d,2}\beta_{a,d} + X_{d,1}X_{a,2}\beta_{d,a} + X_{d,1}X_{d,2}\beta_{d,d})$$

- To provide some intuition concerning what each of these are capturing, consider the values that each of the genotypes would take for dummy variable Xd,1:

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	-1	-1	-1
A_1A_2	1	1	1
A_2A_2	-1	-1	-1

Modeling epistasis V

$$Y = \gamma^{-1}(\beta_\mu + X_{a,1}\beta_{a,1} + X_{d,1}\beta_{d,1} + X_{a,2}\beta_{a,2} + X_{d,2}\beta_{d,2} + X_{a,1}X_{a,2}\beta_{a,a} + X_{a,1}X_{d,2}\beta_{a,d} + X_{d,1}X_{a,2}\beta_{d,a} + X_{d,1}X_{d,2}\beta_{d,d})$$

- To provide some intuition concerning what each of these are capturing, consider the values that each of the genotypes would take for dummy variable Xa,1,Xa,2:

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	**1**	**0**	**-1**
A_1A_2	**0**	**0**	**0**
A_2A_2	**-1**	**0**	**1**

Modeling epistasis VI

$$Y = \gamma^{-1}(\beta_\mu + X_{a,1}\beta_{a,1} + X_{d,1}\beta_{d,1} + X_{a,2}\beta_{a,2} + X_{d,2}\beta_{d,2} + X_{a,1}X_{a,2}\beta_{a,a} + X_{a,1}X_{d,2}\beta_{a,d} + X_{d,1}X_{a,2}\beta_{d,a} + X_{d,1}X_{d,2}\beta_{d,d})$$

- To provide some intuition concerning what each of these are capturing, consider the values that each of the genotypes would take for dummy variable Xa,1Xd,2 (similarly for Xa,2Xd,1):

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	1	-1	1
A_1A_2	0	0	0
A_2A_2	-1	1	-1

Modeling epistasis VII

$$Y = \gamma^{-1}(\beta_\mu + X_{a,1}\beta_{a,1} + X_{d,1}\beta_{d,1} + X_{a,2}\beta_{a,2} + X_{d,2}\beta_{d,2} + X_{a,1}X_{a,2}\beta_{a,a} + X_{a,1}X_{d,2}\beta_{a,d} + X_{d,1}X_{a,2}\beta_{d,a} + X_{d,1}X_{d,2}\beta_{d,d})$$

- To provide some intuition concerning what each of these are capturing, consider the values that each of the genotypes would take for dummy variable Xd,1,Xd,2:

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	1	-1	1
A_1A_2	-1	1	-1
A_2A_2	1	-1	1

Inference for epistasis I

- To infer epistatic relationships we will use the exact same genetic framework and statistical framework that we have been considering
- For the genetic framework, we are still testing markers that we are assuming are in LD with causal polymorphisms that could have an epistatic relationship (so we are indirectly inferring that there is epistasis from the marker genotypes)
- For inference, we going to estimate epistatic parameters using the same approach as before (!!), i.e. for a linear model:

$$\mathbf{X} = [\mathbf{1}, \mathbf{X}_{a,1}, \mathbf{X}_{d,1}, \mathbf{X}_{a,2}, \mathbf{X}_{d,2}, \mathbf{X}_{a,a}, \mathbf{X}_{a,d}, \mathbf{X}_{d,a}, \mathbf{X}_{d,d}]$$

$$\beta = [\beta_{\mu}, \beta_{a,1}, \beta_{d,1}, \beta_{a,2}, \beta_{d,2}, \beta_{a,a}, \beta_{a,d}, \beta_{d,a}, \beta_{d,d}]^{\mathrm{T}}$$

$$\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Inference for epistasis II

- For hypothesis testing, we will just use an LRT calculated the same way as before (!!)
- For an F-statistic for a linear regression and for logistic estimate the parameters under the null and alternative model and substitute these into the likelihood equations that have the same form as before (with some additional dummy variables and parameters)
- The only difference is the degrees of freedom for a given test we consider = number of parameters in the alternative model - the number of parameters in the null model

Inference for epistasis III

- For example, we could use the entire model to test the same hypothesis that we have been considering for a single marker:

$$H_0 : \beta_{a,1} = 0 \cap \beta_{d,1} = 0$$

$$H_A : \beta_{a,1} \neq 0 \cup \beta_{d,1} \neq 0$$

- We could also test whether either marker has evidence of being a causal polymorphism:

$$H_0 : \beta_{a,1} = 0 \cap \beta_{d,1} = 0 \cap \beta_{a,2} = 0 \cap \beta_{d,2} = 0$$

$$H_A : \beta_{a,1} \neq 0 \cup \beta_{d,1} \neq 0 \cup \beta_{a,2} \neq 0 \cup \beta_{d,2} \neq 0$$

- We can also test just for epistasis (note this is equivalent to testing an interaction effect in an ANOVA!):

$$H_0 : \beta_{a,a} = 0 \cap \beta_{a,d} = 0 \cap \beta_{d,a} = 0 \cap \beta_{d,d} = 0$$

$$H_A : \beta_{a,a} \neq 0 \cup \beta_{a,d} \neq 0 \cup \beta_{d,a} \neq 0 \cup \beta_{d,d} \neq 0$$

- We can also test the entire model (what is the interpretation in this case!?):

$$H_0 : \beta_{a,1} = 0 \cap \beta_{d,1} = 0 \cap \beta_{a,2} = 0 \cap \beta_{d,2} = 0 \cap \beta_{a,a} = 0 \cap \beta_{a,d} = 0 \cap \beta_{d,a} = 0 \cap \beta_{d,d} = 0$$

$$H_A : \beta_{a,1} \neq 0 \cup \beta_{d,1} \neq 0 \cup \beta_{a,2} \neq 0 \cup \beta_{d,2} \neq 0 \cup \beta_{a,a} \neq 0 \cup \beta_{a,d} \neq 0 \cup \beta_{d,a} \neq 0 \cup \beta_{d,d} \neq 0$$

Analysis with more phenotypes

- So far, we have considered a GWAS analysis where we have a single phenotype and many genotypes, the latter collected by genomics technologies
- Genomics technologies can also be used to measure many phenotypes (e.g., genome-wide gene expression, proteomics, etc.)
- We also often have a situation where we have both many genotypes and many phenotypes
- The framework you have learned in this class still applies (!!), i.e., the first step in these analyses is still testing pairs of variables at a time

Many phenotypes and one experimental condition I

- Consider a case where you have collected genome-wide gene expression or proteomic data for a tissue of a mouse experiment where there are only two conditions: “wild type" and “mutant”:

$$Data = \left[\begin{array}{ccccccc} z_{11} & \cdots & z_{1k} & y_{11} & \cdots & y_{1m} & x_{11} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{n1} & \cdots & z_{nk} & y_{n1} & \cdots & y_{nm} & x_{11} \end{array}\right]$$

- To analyze these data, regress each phenotype (e.g., a gene expression measurement) on the condition (e.g., coded 0 / 1) one phenotype variable at a time (just like a GWAS!!)

Many phenotypes and one experimental condition IV

- From the statistical modeling point of view, we can view a GWAS as a multiple regression model (i.e., a single Y with many X's):

$$Data = \begin{bmatrix} z_{11} & \dots & z_{1k} & y_{1} & & x_{11} & \dots & x_{1N} \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ z_{n1} & \dots & z_{nk} & y_{n} & & x_{11} & \dots & x_{nN} \end{bmatrix}$$

- While for a case with many phenotypes and a single treatment (e.g., a single genotype) the correct model is a multivariate regression (i.e., many Y's with a single X)

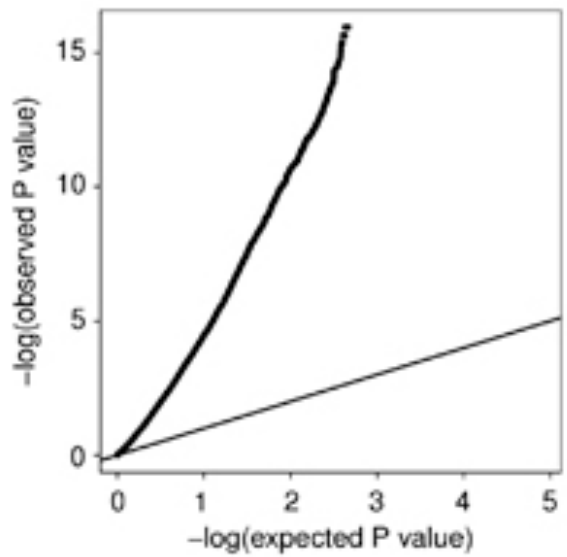
$$Data = \left[\begin{array}{cccccc|c} z_{11} & \dots & z_{1k} & y_{11} & \dots & y_{1m} & x_{11} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{n1} & \dots & z_{nk} & y_{n1} & \dots & y_{nm} & x_{11} \end{array}\right]$$

- We could also have many phenotypes and many genotypes (e.g., eQTL)

$$Data = \begin{bmatrix} z_{11} & \dots & z_{1k} & y_{11} & \dots & y_{1m} & x_{11} & \dots & x_{1N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{n1} & \dots & z_{nk} & y_{n1} & \dots & y_{nm} & x_{11} & \dots & x_{nN} \end{bmatrix}$$

Many phenotypes and one experimental condition II

- There is one important diagnostic difference in the many phenotype analysis: your QQ plots need not conform to the rules of GWAS QQ plots (please take note of this!!)



- That is, when you have a single treatment (or genotype) where you are considering the impact on many phenotypes, it is possible the treatment / genotype impacts many phenotypes (and therefore produces many significant tests!)

Many phenotypes and one experimental condition III

- Why is this?
- That is, why is it that when analyzing GWAS data (=regressing one phenotype on many genotypes) the correct statistical model fitting cannot produce many highly significant tests while an analysis of many phenotypes on one genotype can produce many significant test results (and be the appropriate test result)
- The reason is in a GWAS, we are assuming the underlying true case is many causal genotypes each contributing to variation in the one phenotype, such that if there are many, each of their effects is relatively small (!!)
- In a many phenotypes with one treatment situation, the treatment (or genotype) many separately impact many of the phenotypes (!!)

Multiple and multivariate models I

- While the right first analysis step when dealing with many variables is testing pairs of variables at a time (e.g., one phenotype - one genotype) could we construct statistical models that consider more genotypes or more phenotypes at the same time?
- Yes!
- We could fit multiple regressions with many genotypes (you've done multiple regressions already!)
- We could fit multivariate regressions with many Y's and one treatment
- We could even fit a multivariate-multiple regression model (!!)

Multiple and multivariate models II

- The problem with the multivariate regression approach is many aspects get more complicated and in practice, you often you get the same information as fitting one Y and X pair at a time
- The problem with multiple regressions with many X's is the over-fitting problem, requiring other techniques (e.g., penalized or regularized regressions) and in practice you often get the same information as fitting one Y and X pair
- Same for multivariate-multiple regression situations like eQTL designs (let's take a quick look at this concept first)
- For multiple regressions, we sometimes like to consider a few more X's to capture "interactions" (=epistasis)

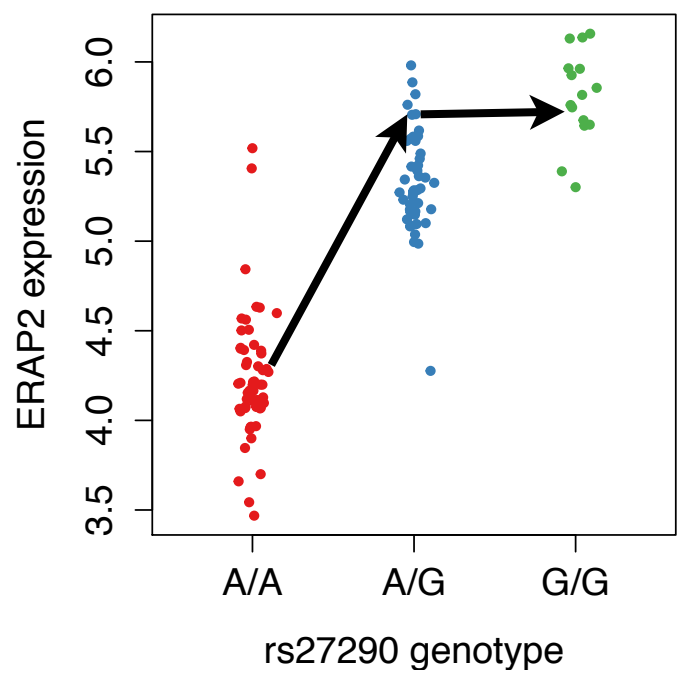
Introduction to eQTL

- **expression Quantitative Trait Locus (eQTL)** - a polymorphic locus where an experimental exchange of one allele for another produces a change in expression on average under specified conditions:

$$A_1 \rightarrow A_2 \Rightarrow \Delta Y | Z$$

- The allelic states defined by the original mutation event define the **causal polymorphism** of the eQTL

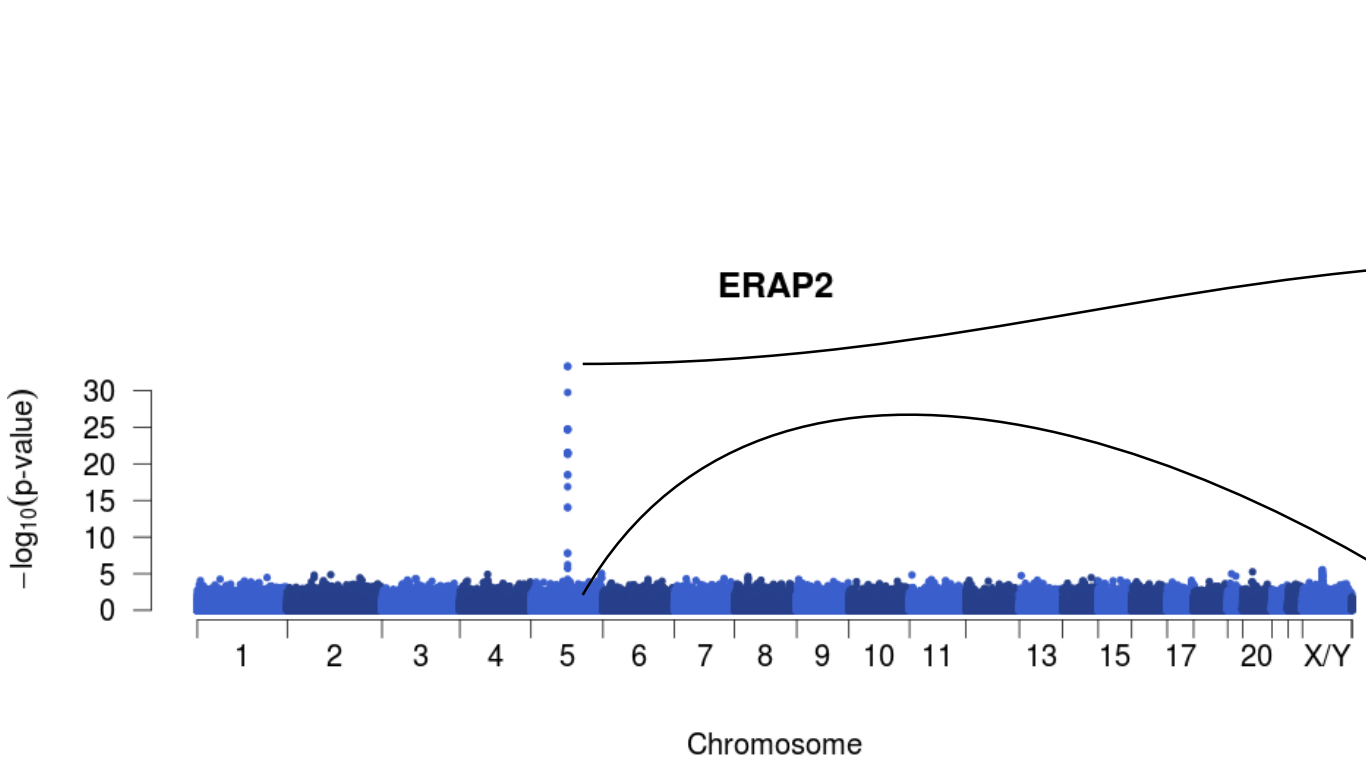
- Intuitive example: if rs27290 was a causal allele, changing A -> G would change the measured expression of ERAP2

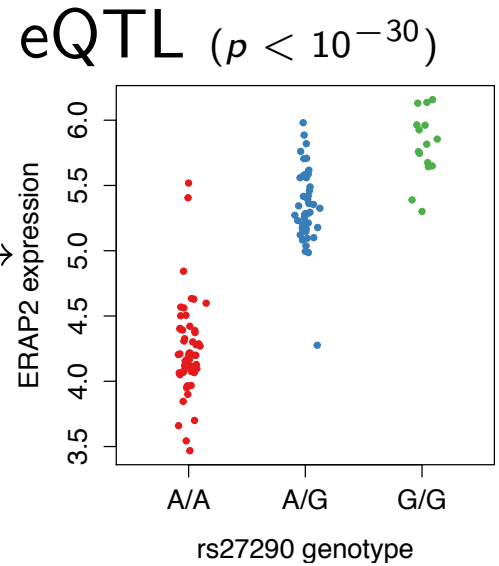


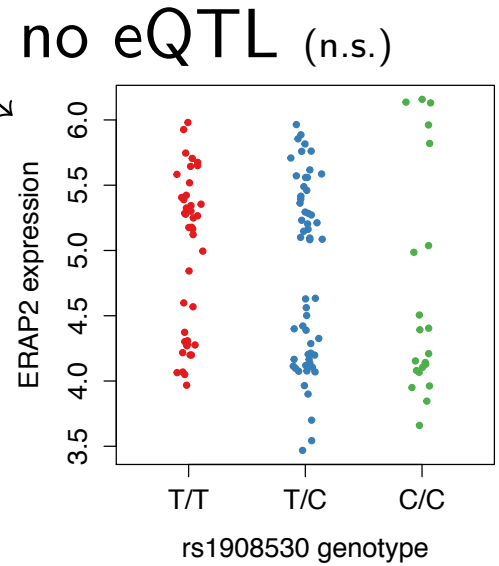
Detecting eQTL from the analysis of genome-wide data

- Since eQTL reflect a case where different allelic combinations (genotypes) lead to different levels of gene expression, we could in theory discover an eQTL by testing for an association between measured genotypes and gene expression levels
- Most eQTL are "discovered" using this type of approach
- A typical (human) eQTL experiment includes *m* (= ~10-30K) expression variables and *N* (= ~0.1-10mil) genotypes measured in *n* individuals sampled from a population
- A typical (most!) analysis of such data proceeds by performing independent statistical tests of (a subset of) genotype-expression pairs, where tests that are significant after a multiple test correct (e.g. Bonferroni), are assumed to indicate an eQTL

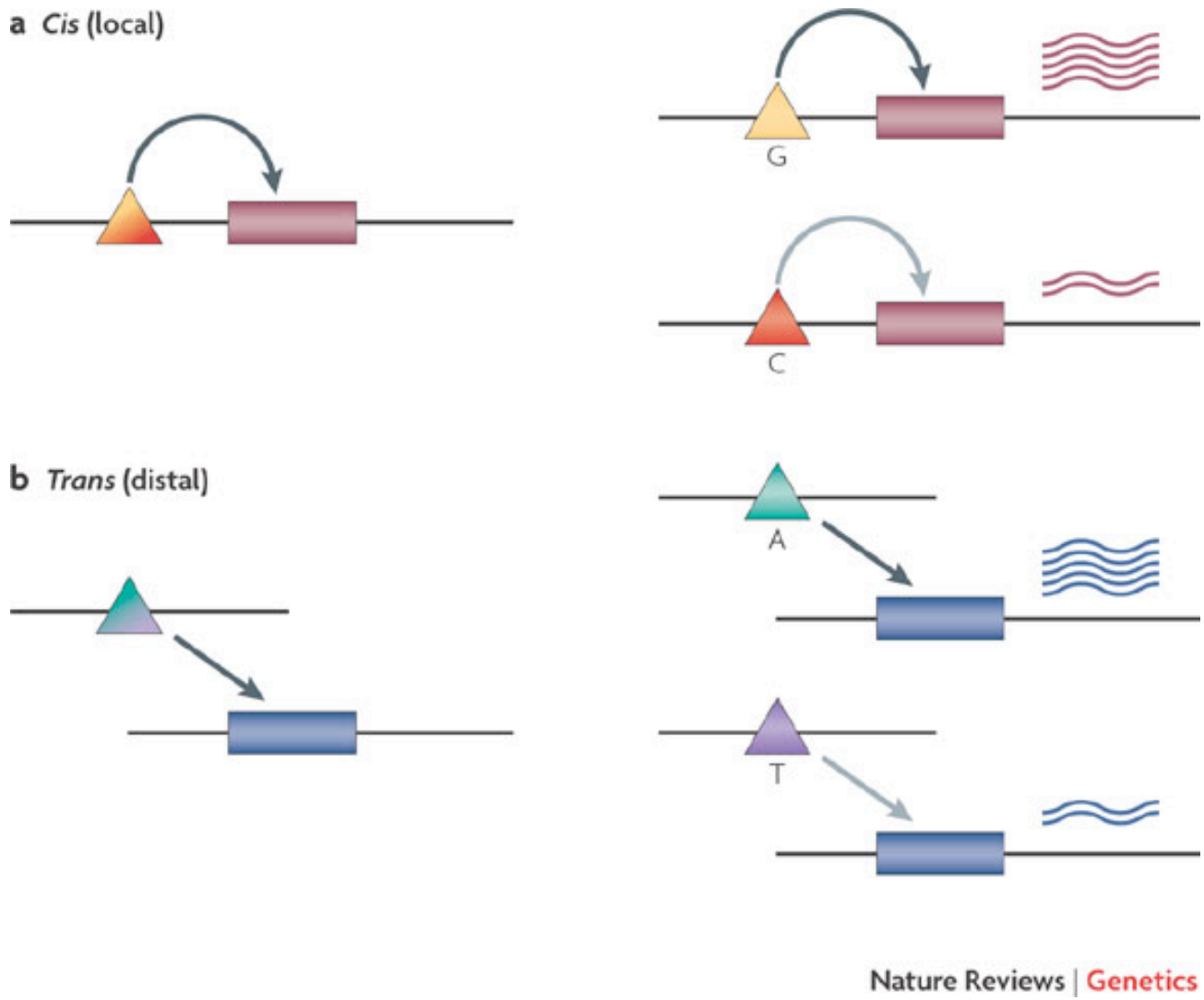
Genome-wide scan for eQTL: typical outcome



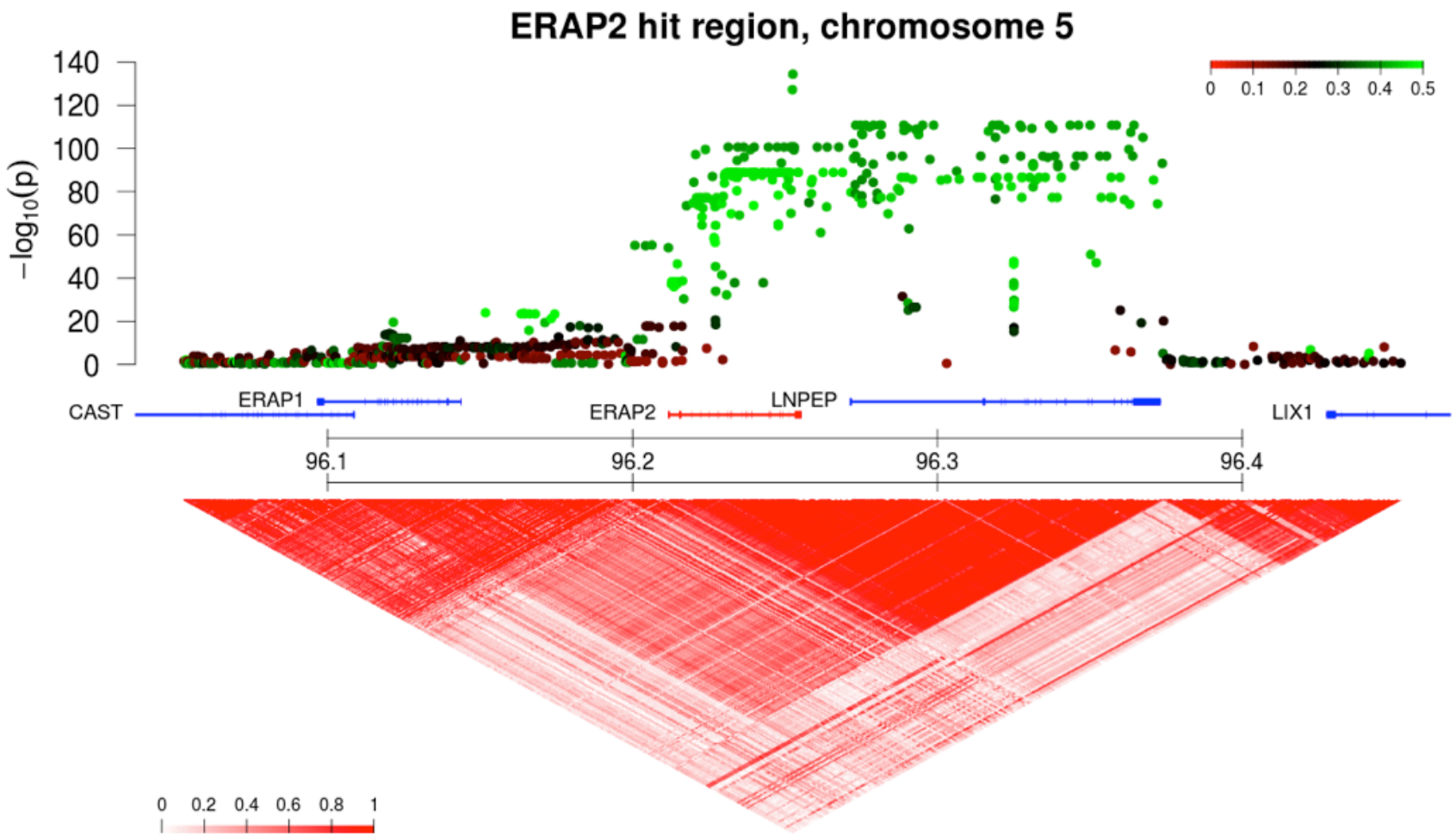




Considering cis- vs trans- eQTL 1



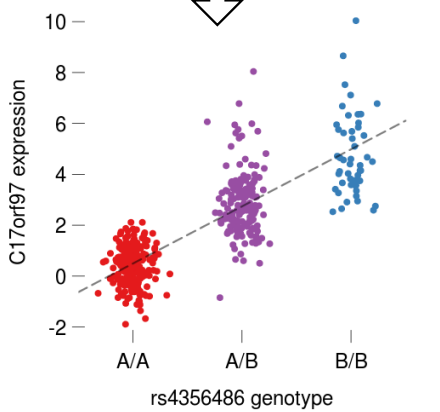
Typical outcome: zooming in and "*cis*-" v "*trans*-"

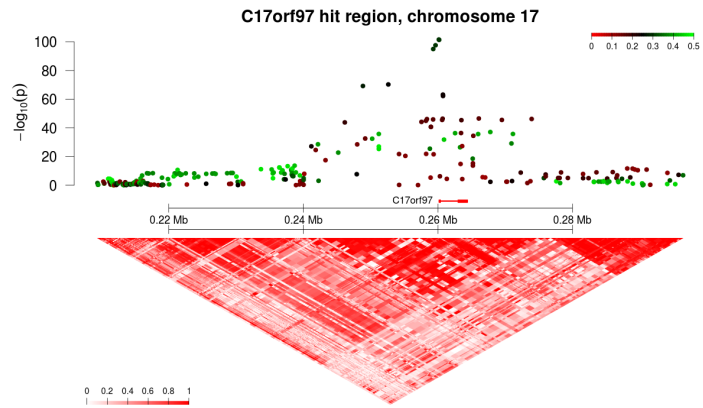


- This is a "cis-"eQTL because the significant genotypes are in the same location as the expressed gene (otherwise, it would be a "trans-"eQTL)
- Most eQTL are "cis-", which makes biological sense

Genome-wide identification of eQTL

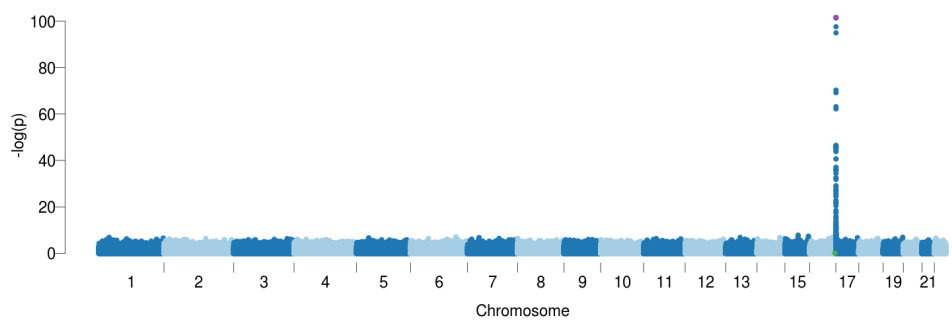
one gene, one SNP

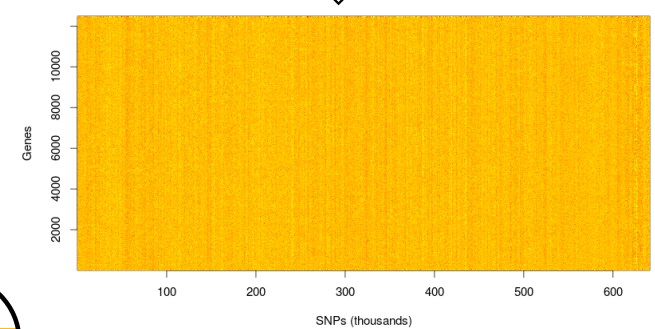




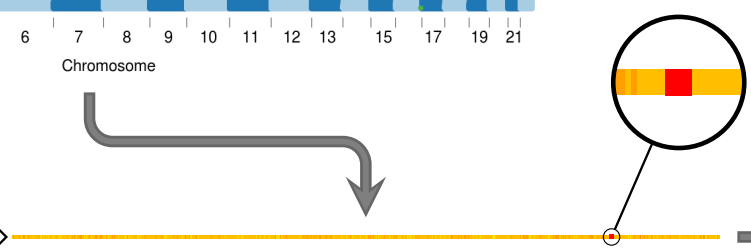
one gene, multiple SNPs

all genes, all SNPs





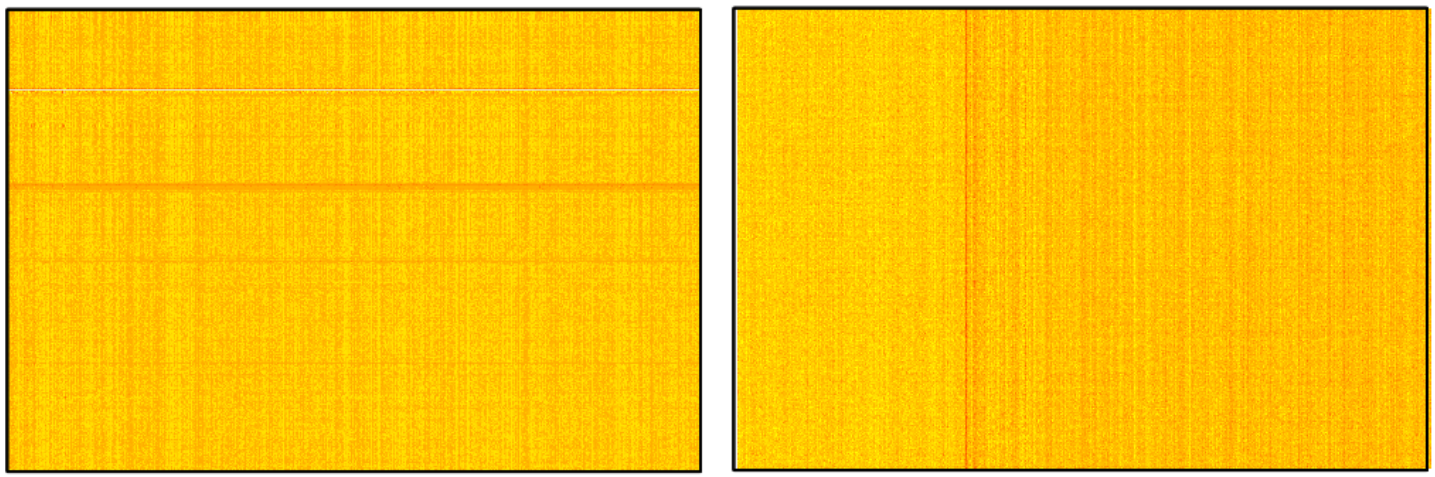
one gene, all SNPs



Advanced Topic: population and hidden factors

- Population structure and hidden factors can cause false positive associations = correlations that don't represent true genetic effects

These effects are visible on the p-value heatmap:



population structure **hidden factor**

- We can sometimes remove these artifacts by including appropriate covariates in our analysis in a mixed model or by using a hidden factor analysis

That's it for today

- See you next time!