Lecture 2: Genetic Association Testing with Quantitative Traits

Instructors: Timothy Thornton and Michael Wu

Summer Institute in Statistical Genetics 2021

Introduction to Quantitative Trait Mapping

- In the previous session, we gave an overview of association testing methods when the trait of interest is binary (e.g. 1/0, affected/unaffected, dead/alive),
- Phenotypes of interest are often quantitative, and in this session we focus on the topic of genetic association testing with quantitative traits.
- The field of **quantitative genetics** is the study of the inheritance of continuously measured traits and their mechanisms.
- Vast amounts of literature on this topic!

Introduction to Quantitative Trait Mapping

- Quantitative trait loci (QTL) mapping involves identifying genetic loci that influence the phenotypic variation of a quantitative trait.
- QTL mapping is commonly conducted with GWAS using common variants, such as variants with minor allele frequencies $\geq 1\% - 5\%$
- There generally is no simple Mendelian basis for variation of quantitative traits
- Some quantitative traits can be largely influenced by a single gene as well as by environmental factors

Introduction to Quantitative Trait Mapping

- Influences on a quantitative trait can also be due to a number of genes with similar (or differing) effects
- Many quantitative traits of interest are complex where phenotypic variation is due to a combination of both multiple genes and environmental factors
- Examples: Blood pressure, cholesterol levels, IQ, height, weight, etc.

Quantitative Genetic Model

- The classical quantitative genetics model introduced by Ronald Fisher (1918) is $Y = G + E$, where Y is the phenotypic value, G is the genetic value, and E is the environmental deviation.
- G is the combination of all genetic loci that influence the phenotypic value and E consists of all non-genetic factors that influence the phenotype
- The mean environmental deviation E is generally taken to be 0 so that the mean genotypic value is equal to the mean phenotypic value, i.e., $E(Y) = E(G)$

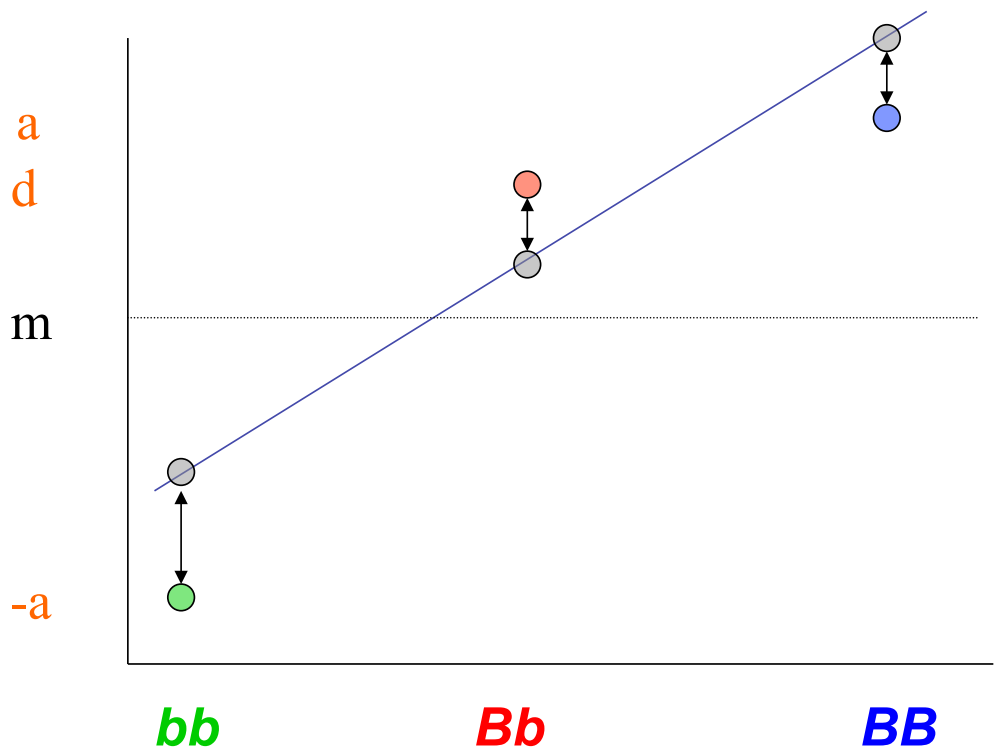
Quantitative Genetic Model

- Consider a single locus. Fisher modeled the genotypic value G with a linear regression model (least squares) where the genotypic value can be partitioned into an additive component (A) and deviations from additivity as a result of dominance (D), where

$$G = A + D$$

Linear Regression Model for Genetic Values

Falconer model for single biallelic QTL



Var (X) = Regression Variance + Residual Variance
= Additive Variance + Dominance Variance

Components of Genetic Variance

- From the properties of least squares, the residuals are orthogonal to the fitted values, and thus $Cov(A, D) = 0$. So we have that

$$Var(G) = Var(A) + Var(D)$$

or

$$\sigma^2_G = \sigma^2_A + \sigma^2_D$$

- σ^2_A is the **additive genetic variance**. It is the genetic variance associated with the average additive effects of alleles
- σ^2_D is the **dominance genetic variance**. It is the genetic variance associated with the dominance effects.

Heritability

- The heritability of a trait is written in terms of the components of variances of the trait.
- Remember that $Y = G + E = A + D + E$
- The following ratio of variance components

$$h^2 = \frac{\sigma_A^2}{\sigma_Y^2}$$

  is defined to be the **narrow-sense heritability** (or simply heritability)
- h^2 is the proportion of the total phenotypic variance that is due to additive effects.
- Heritability can also be viewed as the extent to which phenotypes are determined by the alleles transmitted from the parents.

Heritability

- The **broad-sense heritability** is defined to be

$$H^2 = \frac{\sigma^2_G}{\sigma^2_Y}$$

- H^2 is the proportion of the total phenotypic variance that is due to all genetic effects (additive and dominance)
- There are a number of methods for heritability estimation of a trait.

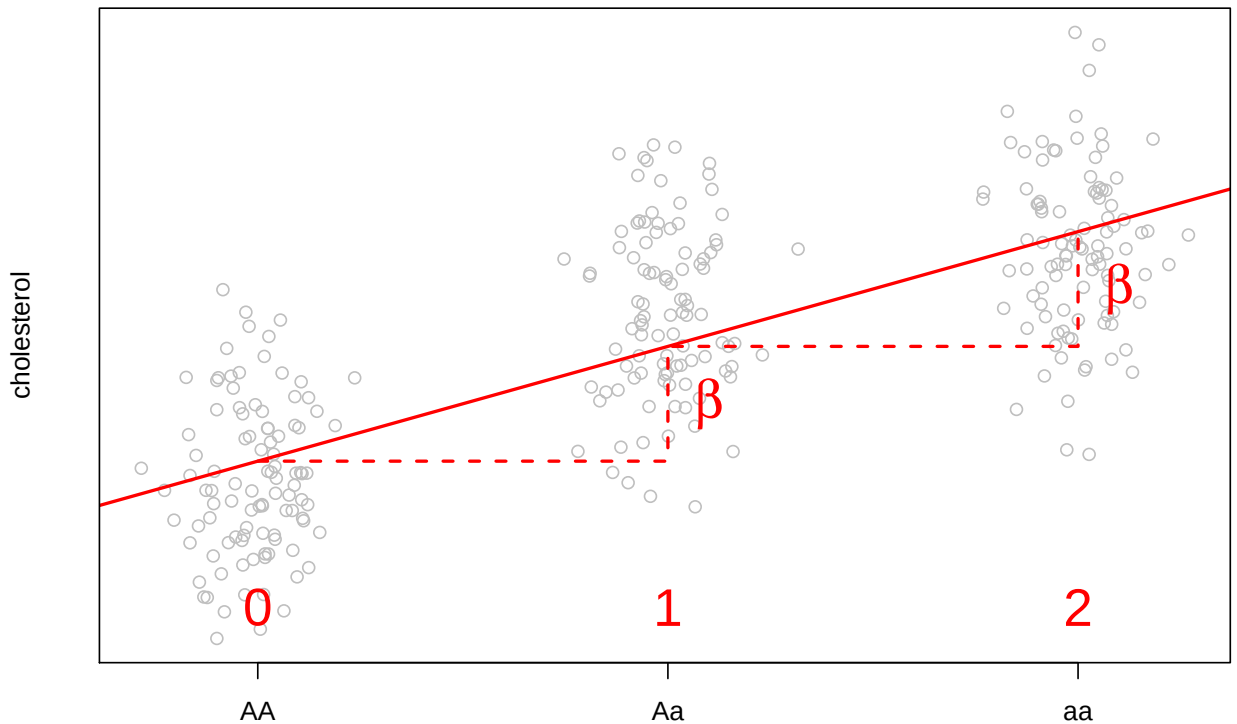
QTL Mapping

- For traits that are heritable, i.e., traits with a non-negligible genetic component that contributes to phenotypic variability, identifying (or mapping) QLT that influence the trait is often of interest.
- Linear regression models are commonly used for QTL mapping
- Linear regression models will often include a single genetic marker (e.g., a SNP) as predictor in the model, in addition to other relevant covariates (such as age, sex, etc.), with the quantitative phenotype as the response

Linear regression with SNPs

Many analyses fit the 'additive model'

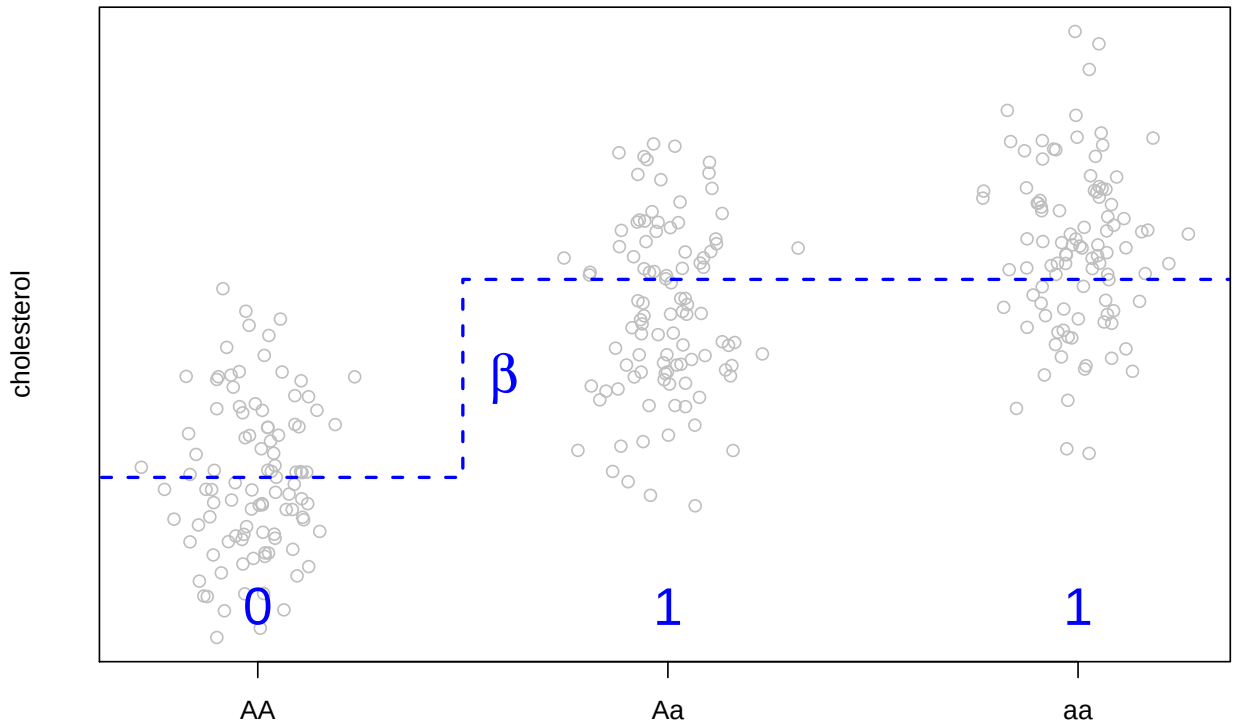
$$y = \beta_0 + \beta \times \#\text{minor alleles}$$



Linear regression, with SNPs

An alternative is the 'dominant model';

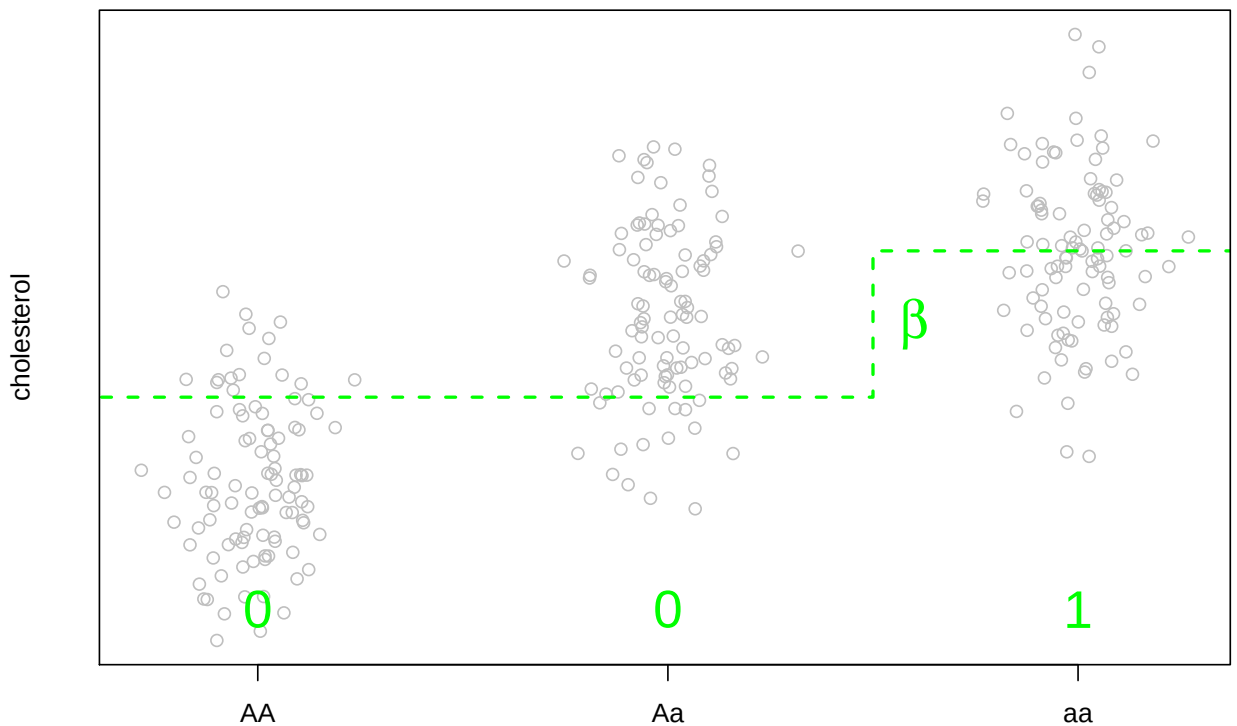
$$y = \beta_0 + \beta \times (G \neq AA)$$



Linear regression, with SNPs

or the 'recessive model';

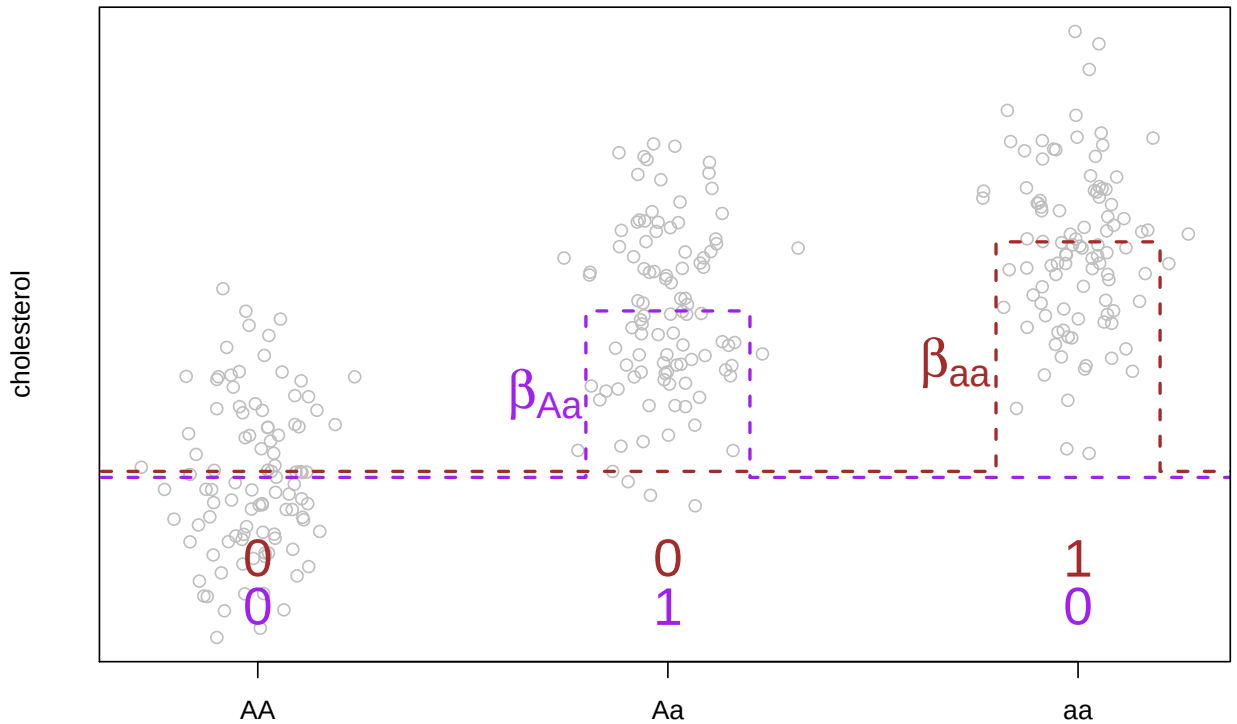
$$y = \beta_0 + \beta \times (G == aa)$$



Linear regression, with SNPs

Finally, the 'two degrees of freedom model';

$$y = \beta_0 + \beta_{Aa} \times (G == Aa) + \beta_{aa} \times (G == aa)$$

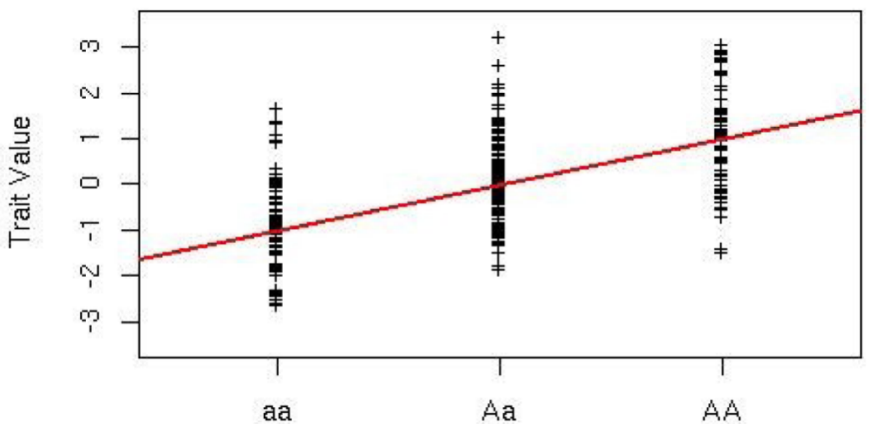


Additive Genetic Model

- Most GWAS perform single SNP association testing with linear regression assuming an additive model.

Unrelated Samples

$$\hat{y}_i = \mu + \hat{\beta} x_i$$



Additive Genetic Model

- The additive linear regression model also has a nice interpretation, as we saw from Fisher's classical quantitative trait model!
- The coefficient of determination (r^2) of an additive linear regression model gives an estimate of the proportion of phenotypic variation that is explained by the SNP (or SNPs) in the model, e.g., the "SNP heritability"

Additive Genetic Model

- Consider the following additive model for association testing with a quantitative trait and a SNP with alleles A and a:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

  where X is the number of copies of the reference allele A.
- What would your interpretation of ϵ be for this particular model?

Association Testing with Additive Model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Two test statistics for $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$

$$T = \frac{\hat{\beta}_1}{\sqrt{var(\hat{\beta}_1)}} \sim t_{N-2} \approx N(0,1) \text{ for large } N$$

$$T^2 = \frac{\hat{\beta}_1^2}{var(\hat{\beta}_1)} \sim F_{1,N-2} \approx \chi_1^2 \text{ for large } N$$

where

$$var(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{S_{XX}}$$

and S_{XX} is the corrected sum of squares for the X_i's

Statistical Power for Detecting QTL

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- We can also calculate the power for detecting a QTL for a given effect size β_1 for a SNP.
- For simplicity, assume that Y has been a standardized so that with $\sigma^2_Y = 1$.
- Let p be the frequency of the A allele in the population

$$\sigma^2_Y = \beta_1^2 \sigma^2_X + \sigma^2_\epsilon = 2p(1-p)\beta_1^2 + \sigma^2_\epsilon$$

- Let $h^2_s = 2p(1-p)\beta_1^2$, so we have $\sigma^2_Y = h^2_s + \sigma^2_\epsilon$
- Interpret h^2_s (note that we assume that trait is standardized such that $\sigma^2_Y = 1$)

Statistical Power for Detecting QTL

- Also note that $\sigma_\epsilon^2 = 1 - h_s^2$, so we can write $Var(\hat{\beta}_1)$ as the following:

$$var(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{S_{XX}} \approx \frac{\sigma_\epsilon^2}{N\,(2p(1-p))} = \frac{1 - h_s^2}{2Np(1-p)}$$

- To calculate power of the test statistic T^2 for a given sample size N, we need to first obtain the expected value of the non-centrality parameter λ of the chi-squared (χ^2) distribution which is the expected value of the test statistic T squared:

$$\lambda = [E(T)]^2 \approx \frac{\beta_1^2}{var(\hat{\beta}_1)} = \frac{Nh_s^2}{1 - h_s^2}$$

since $h_s^2 = 2p(1-p)\beta_1^2$

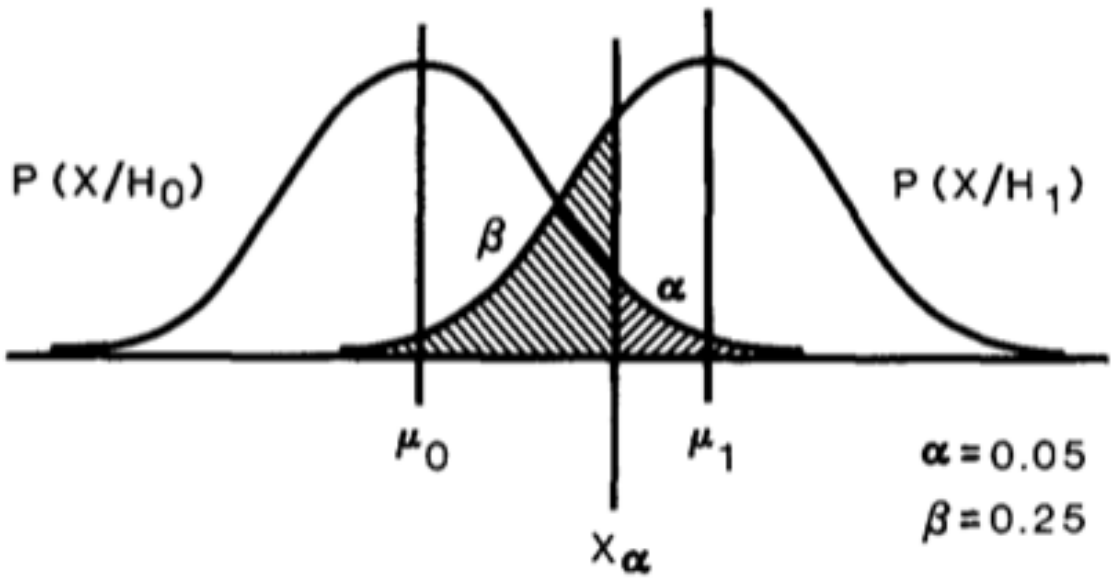
Required Sample Size for Power

- Can also obtain the required sample size given type-I error α and power $1-\beta$, where the type–II error is β :

$$N = \frac{1-h_s^2}{h_s^2}\left(z_{(1-\alpha/2)} + z_{(1-\beta)}\right)^2$$

where $z_{(1-\alpha/2)}$ and $z_{(1-\beta)}$ are the $(1-\alpha/2)$th and $(1-\beta)$th quantiles, respectively, for the standard normal distribution.

Statistical Power for Detecting QTL



Genetic Power Calculator (PGC)
http://pngu.mgh.harvard.edu/~purcell/gpc/

Genetic Power Calculator

Genetic Power Calculator

S. Purcell & P. Sham, 2001-2009

This site provides automated power analysis for variance components (VC) quantitative trait locus (QTL) linkage and association tests in sibships, and other common tests. Suggestions, comments, etc to *Shaun Purcell*.

If you use this site, please reference the following Bioinformatics article:

```
Purcell S, Cherny SS, Sham PC. (2003) Genetic Power Calculator:
design of linkage and association genetic mapping studies of complex
traits. Bioinformatics, 19(1):149-150.
```

Modules

Case-control for discrete traits	Notes
Case-control for threshold-selected quantitative traits	Notes
QTL association for sibships and singletons	Notes
TDT for discrete traits	Notes
TDT and parenTDT with ascertainment	Notes
TDT for threshold-selected quantitative traits	Notes
Epistasis power calculator	Notes
QTL linkage for sibships	Notes
Probability Function Calculator	Notes

Genetic Power Calculator

QTL Association for Sibships

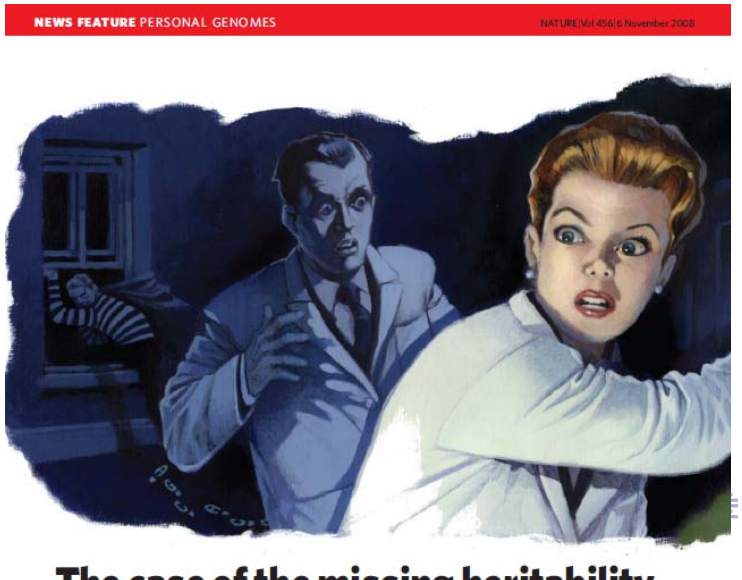
```
Total QTL variance                 :          (0 - 1)
Dominance : additive QTL effects   :          (0 - 1) ☑ No dominance (* see below)
QTL increaser allele frequency     :          (0 - 1)
Marker M1 allele frequency         :          (0 - 1)
Linkage disequilibrium (D-prime)   :          (0 - 1)
Sibling correlation                :          (0 - 1) (* see below)

Sample Size                        :          (0 - 10000000) (N=families, not individuals)
Sibship Size                       :  Pairs   ☑ Both parents genotyped

User-defined type I error rate     :  0.05    (0.00000001 - 0.5)
User-defined power: determine N    :  0.80    (0 - 1)
(1 - type II error rate)
```

Process Reset

Missing Heritability

Disease	Number of loci	Percent of Heritability Measure Explained	Heritability Measure
Age-related macular degeneration	5	50%	Sibling recurrence risk
Crohn's disease	32	20%	Genetic risk (liability)
Systemic lupus erythematosus	6	15%	Sibling recurrence risk
Type 2 diabetes	18	6%	Sibling recurrence risk
HDL cholesterol	7	5.2%	Phenotypic variance
Height	40	5%	Phenotypic variance
Early onset myocardial infarction	9	2.8%	Phenotypic variance
Fasting glucose	4	1.5%	Phenotypic variance

- GWAS works
- Effect sizes are typically small
  - Disease: OR ~1.1 to ~1.3
  - Quantitative traits: % var explained <<1%



The case of the missing heritability

Genetic Power Calculator (Shaun Purcell)

http://pngu.mgh.harvard.edu/~purcell/gpc/

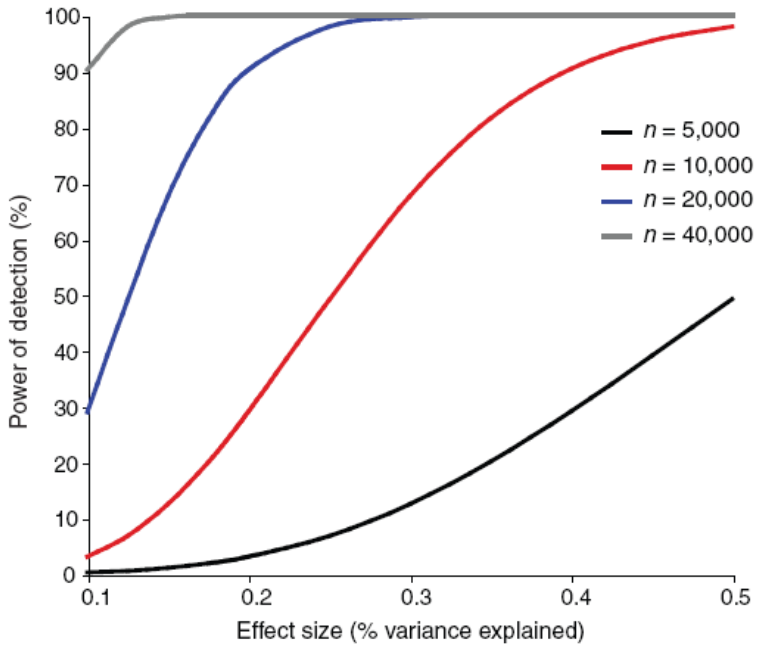
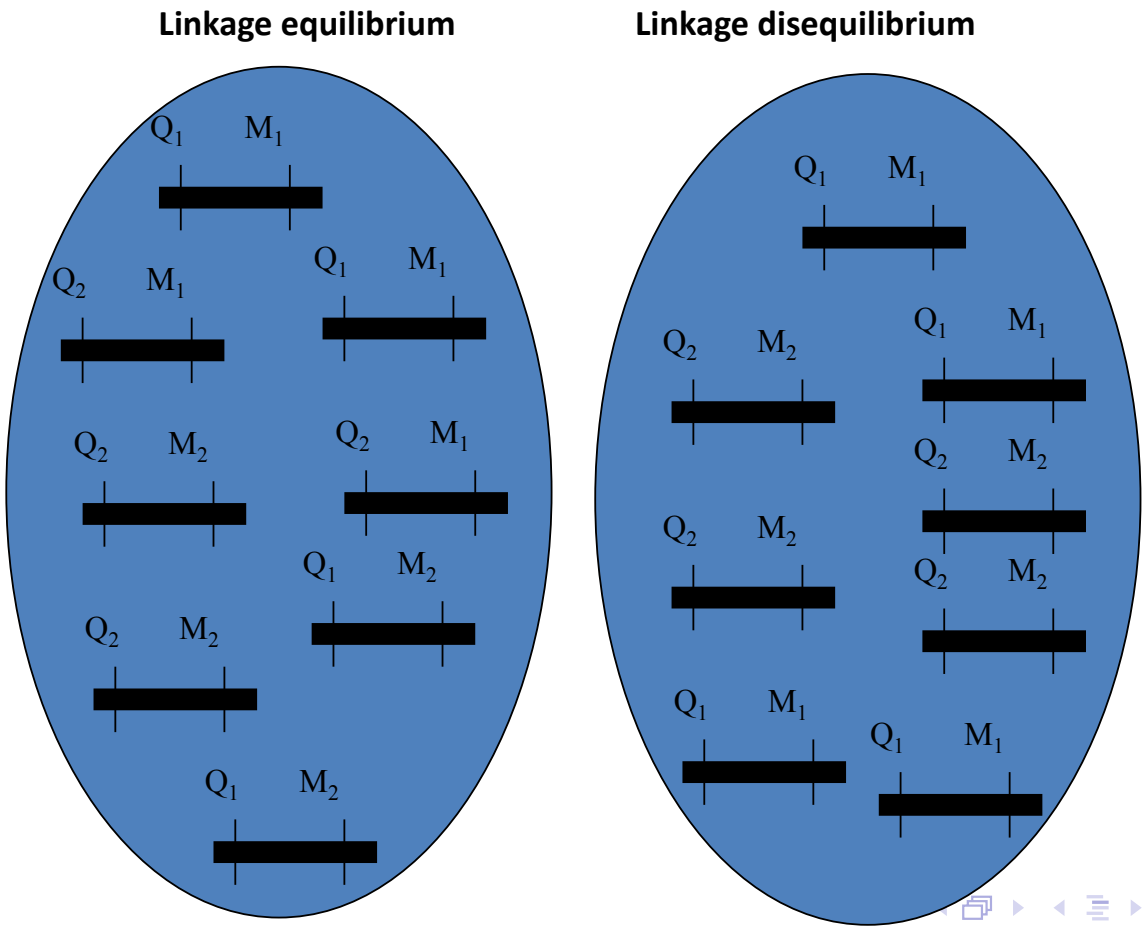


Figure 1 Statistical power of detection in GWAS for variants that explain 0.1–0.5% of the variation at a type I error rate of 5×10^{-7} (calculated using the Genetic Power Calculator[15]). Shown is the power to detect a variant with a given effect size, assuming this type I error rate, which is typical for a GWAS with a sample size of $n = 5{,}000$–$40{,}000$.

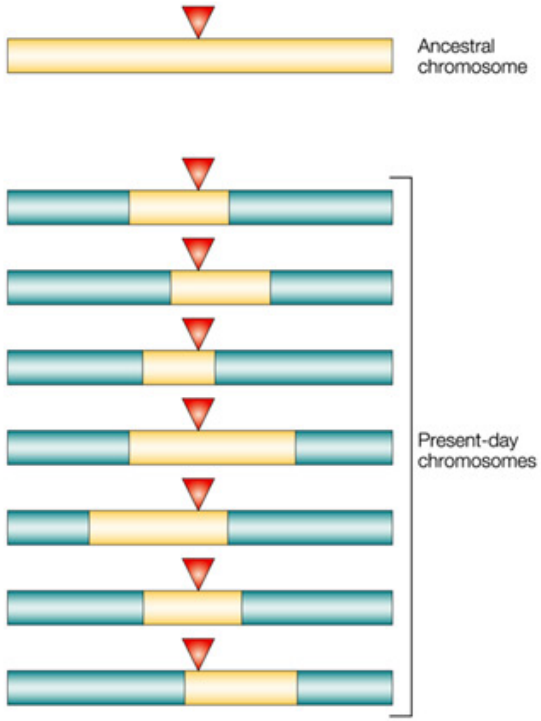
LD Mapping of QTL

- For GWAS, the QTL generally will not be genotyped in a study



LD Mapping of QTL

Linkage disequilibrium around an ancestral mutation



LD Mapping of QTL

- r^2 = LD correlation between QTL and genotyped SNP
- Proportion of variance of the trait explained at a SNP $\approx r^2 h_s^2$
- Required sample size for detection is

$$N \approx \frac{1 - r^2 h_s^2}{r^2 h_s^2} \left(z_{(1-\alpha/2)} + z_{(1-\beta)} \right)^2$$

- Power of LD mapping depends on the experimental sample size, variance explained by the causal variant and LD with a genotyped SNP