QTL Association Mapping

Introduction to Quantitative Trait Mapping

- We previously focused on obtaining variance components of a quantitative trait to determine the proportion of the variance of the trait that can be attributed to both genetic (additive and dominance) and environment (shared and unique) factors
- We demonstrated that resemblance of trait values among relatives we can be used to obtain estimates of the variance components of a quantitative trait without using genotype data.
- Quantitative trait loci (QTL) mapping involves identifying genetic loci that influence the variation of a quantitative trait.

Introduction to Quantitative Trait Mapping

- 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
- Influences on a quantitative trait can be due to a a large 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.

Partition of Phenotypic Values

- Today we will focus on
  - QTL association mapping
  - Contribution of a QTL to the variance of a quantitative trait
  - Statistical power for detecting QTL in GWAS
- Consider once again the classical quantitative genetics model of $Y = G + E$ where Y is the phenotype value, G is the genotypic value, and E is the environmental deviation that is assumed to have a mean of 0 such that $E(Y) = E(G)$

Representation of Genotypic Values

- For a single locus with alleles A_1 and A_2, the genotypic values for the three genotypes can be represented as follows

$$\text{Genotype Value} = \begin{cases} -a & \text{if genotype is } A_2A_2 \\ d & \text{if genotype is } A_1A_2 \\ a & \text{if genotype is } A_1A_1 \end{cases}$$

- If p and q are the allele frequencies of the A_1 and A_2 alleles, respectively in the population, we previously showed that

$$\mu_G = a(p-q) + d(2pq)$$

  and that the genotypic value at a locus can be decomposed into additive effects and dominance deviations:

$$G_{ij} = G_{ij}^A + \delta_{ij} = \mu_G + \alpha_i + \alpha_j + \delta_{ij}$$

Decomposition of Genotypic Values

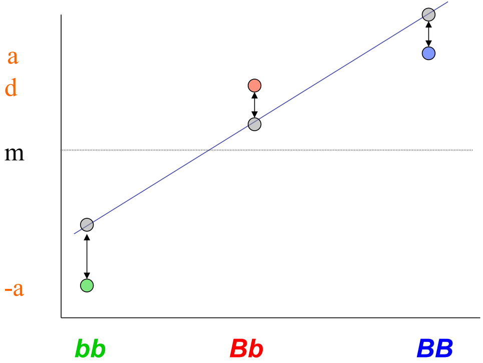
- The model can be given in terms of a linear regression of genotypic values on the number of copies of the A_1 allele such that:

  $$G_{ij} = \beta_0 + \beta_1 X_1^{ij} + \delta_{ij}$$

  where X_1^{ij} is the number of copies of the type A_1 allele in genotype G_{ij}, and with $\beta_0 = \mu_G + 2\alpha_2$ and $\beta_1 = \alpha_1 - \alpha_2 = \alpha$, the average effect of allele substitution.
- Recall that $\alpha = a + d(q - p)$ and that $\alpha_1 = q\alpha$ and $\alpha_2 = -p\alpha$

Linear Regression Figure for Genetic Values

Falconer model for single biallelic QTL



Var (X) = Regression Variance + Residual Variance

= Additive Variance + Dominance Variance

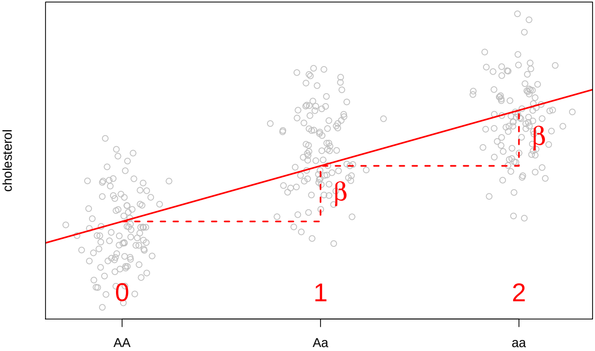
QTL Mapping

- 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.
- Genome-wide association studies (GWAS) are commonly used for the identification of QTL
- Single SNP association testing with linear regression models are often used in GWAS
- 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’

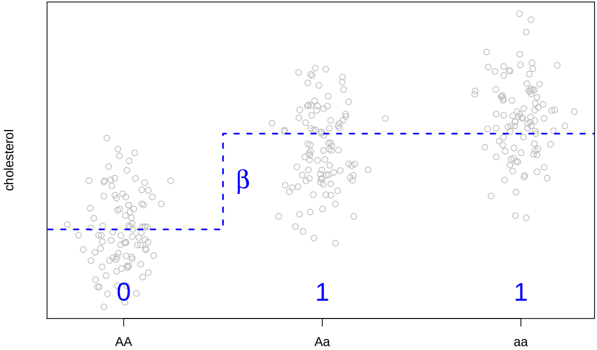
$$y = \beta_0 + \beta \times \#\text{minor alleles}$$



Linear regression, with SNPs

An alternative is the 'dominant model';

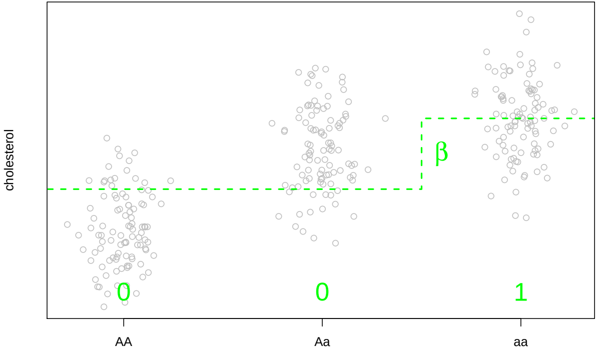
$$y = \beta_0 + \beta \times (G \neq AA)$$



Linear regression, with SNPs

or the ‘recessive model’;

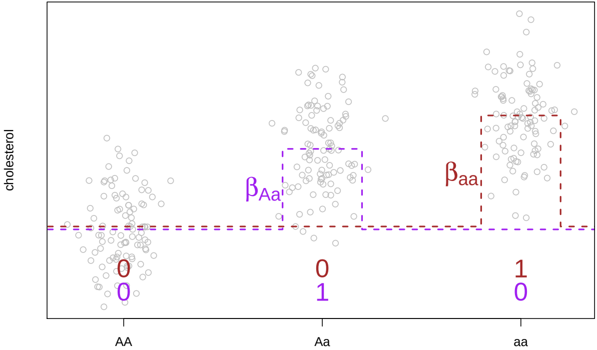
$$y = \beta_0 + \beta \times (G == aa)$$



Linear regression, with SNPs

Finally, the ‘two degrees of freedom model’;

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

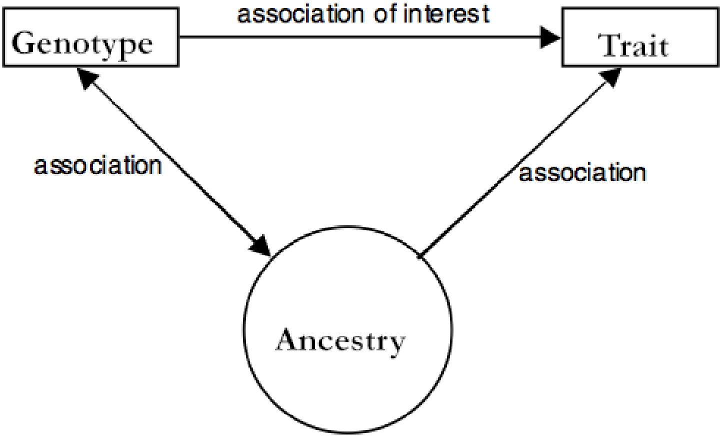


Association Testing with Dependent Samples

- The observations in genetic association studies can have several sources of dependence, including:
  - population structure, i.e., ancestry differences among sample individuals
  - relatedness among the sampled individuals, some of which might be known and some unknown.
- Failure to appropriately account for this structure can invalidate association results that are based on an assumption of independence and population homogeneity.

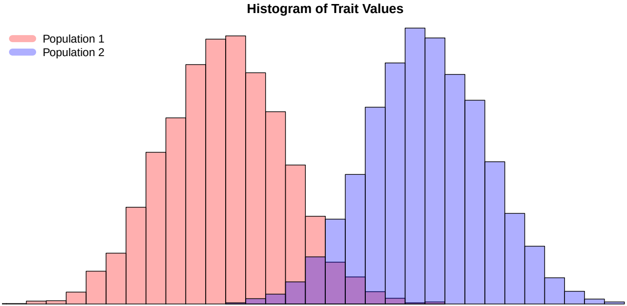
Confounding due to Ancestry

- Ethnic groups (and subgroups) often share distinct dietary habits and other lifestyle characteristics that leads to many traits of interest being correlated with ancestry and/or ethnicity.



Spurious Association

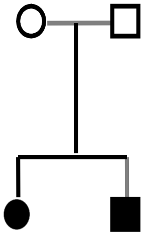
- Quantitative trait association test
  - Test for association between genotype and trait value
- Consider sampling from 2 populations:

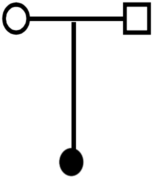


  - Blue population has higher trait values.
  - Different allele frequency in each population
    $\Longrightarrow$ spurious association between trait and genetic marker for samples containing individuals from both populations

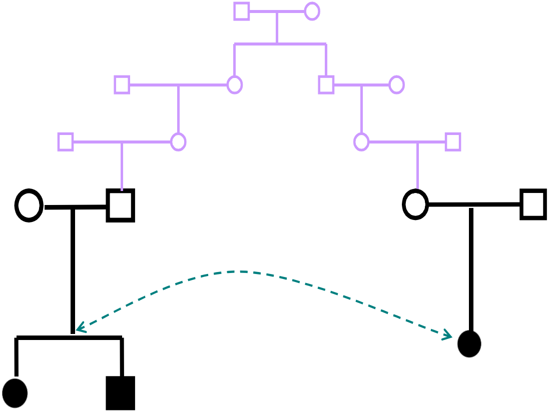
Incomplete Genealogy

- Cryptic and/or misspecified relatedness among the sample individuals can also lead to spurious association in genetic association studies





Incomplete Genealogy



Genotype and Phenotype Data

- Linear mixed models have been demonstrated to be a flexible approach for association testing in samples with population and/or pedigree structure.
- Suppose the data for the genetic association study include genotype and phenotype on a sample of n individuals
- Let $\mathbf{Y} = (Y_1, \ldots Y_n)^T$ denote the $n \times 1$ vector of phenotype data, where Y_i is the quantitative trait value for the ith individual.
- Consider testing SNP s in a genome-screen for association with the phenotype, where $\mathbf{G_s} = (G_1^s, \ldots G_n^s)^T$ is $n \times 1$ vector of the genotypes, where $G_i^s = 0, 1,$ or 2, according to whether individual i has, respectively, 0, 1 or 2 copies of the reference allele at SNP s.

Association Testing with Cryptic Structure

- Consider the following model:

$$\mathbf{Y} = \mathbf{W}\beta + \mathbf{G_s}\gamma + \mathbf{g} + \varepsilon$$

- $\mathbf{W}$ is an $n \times (w+1)$ matrix of relevant covariates that includes an intercept
- β is the $(w+1) \times 1$ vector of covariate effects, including intercept
- γ is the (scalar) association parameter of interest, measuring the effect of genotype on phenotype
- $\mathbf{g}$ is a length n random vector of polygenic effects with $\mathbf{g} \sim N(\mathbf{0}, \sigma_g^2 \mathbf{\Psi})$
- σ_g^2 represents additive genetic variance and $\mathbf{\Psi}$ is a matrix of pairwise measures of genetic relatedness
- ε is a random vector of length n with $\varepsilon \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I})$
- σ_e^2 represents non-genetic variance due to non-genetic effects assumed to be acting independently on individuals

Mixed Linear Models For Cryptic Structure

- The matrix $\mathbf{\Psi}$ will be generally be unknown when there is population structure (ancestry differences) and/or cryptic relatedness in the sample.
- Kang et al. [Nat Genet, 2010] proposed the EMMAX linear mixed model association method that is based on an empirical genetic relatedness matrix (GRM) $\hat{\mathbf{\Psi}}$ calculated using SNPs from across the genome. The $(i,j)th$ entry of the matrix is estimated by

$$\hat{\mathbf{\Psi}}_{ij} = \frac{1}{S}\sum_{s=1}^{S}\frac{(G_i^s - 2\hat{p}_s)(G_j^s - 2\hat{p}_s)}{2\hat{p}_s(1-\hat{p}_s)}$$

where $\hat{p}_s$ is the sample average allele frequency. S will generally need to be quite large, e.g., larger than 100,000, to capture fine-scale structure.

Kang, Hyun Min, et al. (2010) "Variance component model to account for sample structure in genome-wide association studies." Nature genetics 42

EMMAX Mixed Linear Model For Cryptic Structure

- For genetic association testing, the EMMAX mixed-model approach first considers the following model without including any of the SNPs as fixed effects:

$$\mathbf{Y} = \mathbf{W}\beta + \mathbf{g} + \varepsilon \tag{1}$$

- The variance components, σ_g^2 and σ_e^2, are then estimated using either a maximum likelihood or restricted maximum likelihood (REML), with $\mathbf{Cov}(\mathbf{Y})$ set to $\sigma_g^2\hat{\mathbf{\Psi}} + \sigma_e^2\mathbf{I}$ in the likelihood with fixed $\hat{\mathbf{\Psi}}$

EMMAX Mixed Linear Model For Cryptic Structure

- Once the variance components , σ_g^2 and σ_e^2 are then estimated, association testing of SNP s and phenotype is then based on the model
$$\mathbf{Y} = \mathbf{W}\beta + \mathbf{G^s}\gamma + \mathbf{g} + \varepsilon$$
- The EMMAX association statistic is the score statistic for testing the null hypothesis of $\gamma = 0$ using a generalized regression with $Var(\mathbf{Y}) = \mathbf{\Sigma}$ evaluated at $\hat{\mathbf{\Sigma}} = \hat{\sigma}_g^2 \hat{\mathbf{\Psi}} + \hat{\sigma}_e^2 \mathbf{I}$
- EMMAX calculates $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ only once from model (1) to reduce computational burden.

GEMMA Linear Mixed Model For Cryptic Structure

- Zhou and Stephens [2012, Nat Genet] developed a computationally efficient mixed-model approach named GEMMA
- GEMMA is very similar to EMMAX and is essentially based on the same linear mixed-model as EMMAX

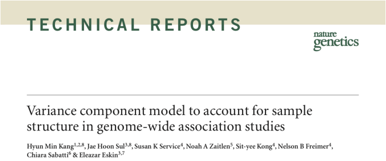
$$\mathbf{Y} = \mathbf{W}\beta + \mathbf{G}^{\mathbf{s}}\gamma + \mathbf{g} + \varepsilon$$

- However, the GEMMA method is an "exact" method that obtains maximum likelihood estimates of variance components $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ for each SNP s being tested for association.

Zhou and Stephens (2012) "Genome-wide efficient mixed-model analysis for association studies" Nature Genetics 44

Linear Mixed Models For Cryptic Structure

- A number of similar linear mixed-effects methods have recently been proposed when there is cryptic structure: Zhang at al. [2010, Nat Genet], Lippert et al. [2011, Nat Methods], Zhou & Stephens [2012, Nat Genet], and Svishcheva [2012, Nat, Genet], and others.

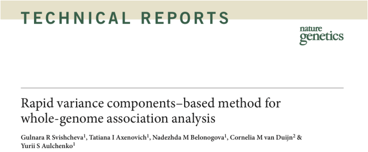


TECHNICAL REPORTS

nature genetics

Variance component model to account for sample structure in genome-wide association studies

Hyun Min Kang[1,2,8], Jae Hoon Sul[3,8], Susan K Service[4], Noah A Zaitlen[5], Sit-yee Kong[4], Nelson B Freimer[4], Chiara Sabatti[6] & Eleazar Eskin[3,7]

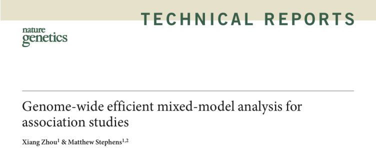


TECHNICAL REPORTS

nature genetics

Rapid variance components–based method for whole-genome association analysis

Gulnara R Svishcheva[1], Tatiana I Axenovich[1], Nadezhda M Belonogova[1], Cornelia M van Duijn[2] & Yurii S Aulchenko[1]

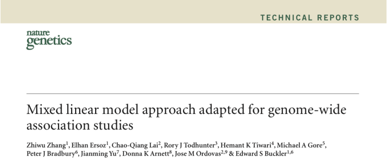


TECHNICAL REPORTS

nature genetics

Genome-wide efficient mixed-model analysis for association studies

Xiang Zhou[1] & Matthew Stephens[1,2]



TECHNICAL REPORTS

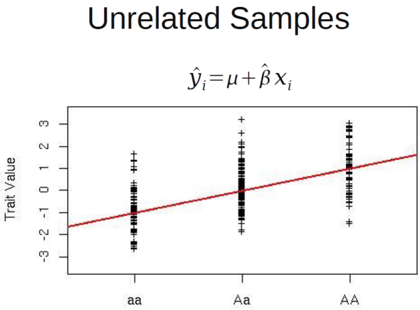
nature genetics

Mixed linear model approach adapted for genome-wide association studies

Zhiwu Zhang[1], Elhan Ersoz[1], Chao-Qiang Lai[2], Rory J Todhunter[3], Hemant K Tiwari[4], Michael A Gore[5], Peter J Bradbury[6], Jianming Yu[7], Donna K Arnett[8], Jose M Ordovas[2,9] & Edward S Buckler[1,6]

Additive Genetic Model

- Most GWAS are performed via single SNP association testing under an additive model.



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:

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

  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

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

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

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

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

where

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

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

Statistical Power for Detecting QTL

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

- 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 $\sigma_Y^2 = 1$.
- Let p be the frequency of the A allele in the population

$$\sigma_Y^2 = \beta_1^2 \sigma_X^2 + \sigma_\varepsilon^2 = 2p(1-p)\beta_1^2 + \sigma_\varepsilon^2$$

- Let $h_s^2 = 2p(1-p)\beta_1^2$, so we have $\sigma_Y^2 = h_s^2 + \sigma_\varepsilon^2$
- Interpret h_s^2 (note that we assume that trait is standardized such that $\sigma_Y^2 = 1$)

Statistical Power for Detecting QTL

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

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

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

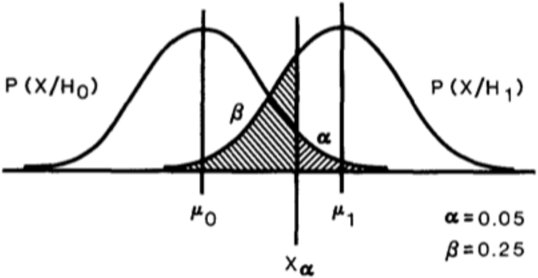
Required Sample Size for Power

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

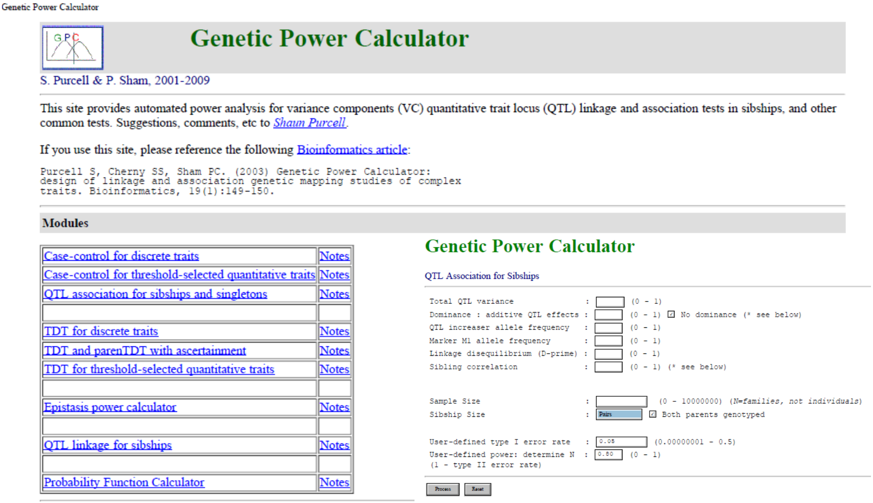
$$N = \frac{1-h_s^2}{h_s^2}\left(z_{(1-\alpha/2)} + z_{(1-\beta)}\right)^2$$

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)
```

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%

NEWS FEATURE PERSONAL GENOMES NATURE|Vol 456|6 November 2008

The case of the missing heritability

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

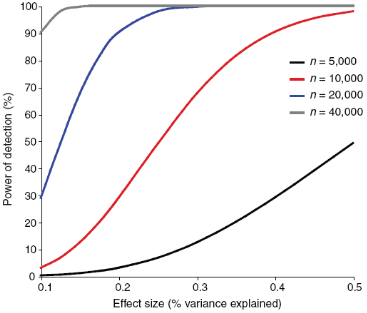
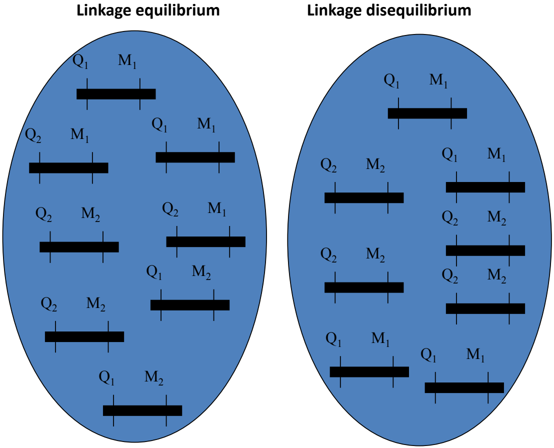


Figure 1 Statistical power of detection in GWAS for variants that explain 0.1–0.5% of the variation at a type I error rate of 5×10^{-7} (calculated using the Genetic Power Calculator[15]). Shown is the power to detect a variant with a given effect size, assuming this type I error rate, which is typical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

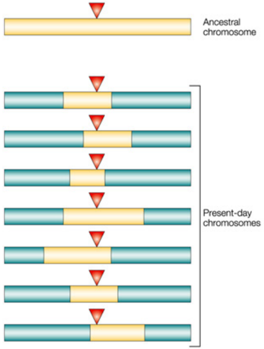
LD Mapping of QTL

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



LD Mapping of QTL

Linkage disequilibrium around an ancestral mutation



Nature Reviews | Genetics

5

[Ardlie et al. 2002]

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