QTL Mapping II: Outbred Populations and Association Mapping

Bruce Walsh lecture notes

Uppsala EQG 2012 course

version 2 Feb 2012

Mapping in Outbred Populations

QTL mapping in outbred populations has far lower power compared to line crosses.

Not every individual is informative for linkage.
an individual must be a double heterozygote
to provide linkage information

Parents can differ in linkage phase, e.g., MQ/mq vs. Mq/mQ. Hence, cannot pool families, rather must analyze each parent separately.

Marker vs. QTL Informative

Can easily check to see if a parent/family is marker informative (at least one parent is a marker heterozygote).

No easy way to check if they are also QTL informative (at least one family is a QTL heterozygote)

A fully-informative parent is both Marker and QTL informative, i.e., a double heterozygote.

Types of families (considering marker information)

Fully Marker informative Family:

$M_iM_j \times M_kM_l$ Both parents different heterozygotes

All offspring are informative in distinguishing alternative alleles from both parents.

Backcross family

$M_iM_j \times M_kM_k$ One parent a marker homozygote

All offspring informative in distinguishing heterozygous parent's alternative alleles

Intercross family

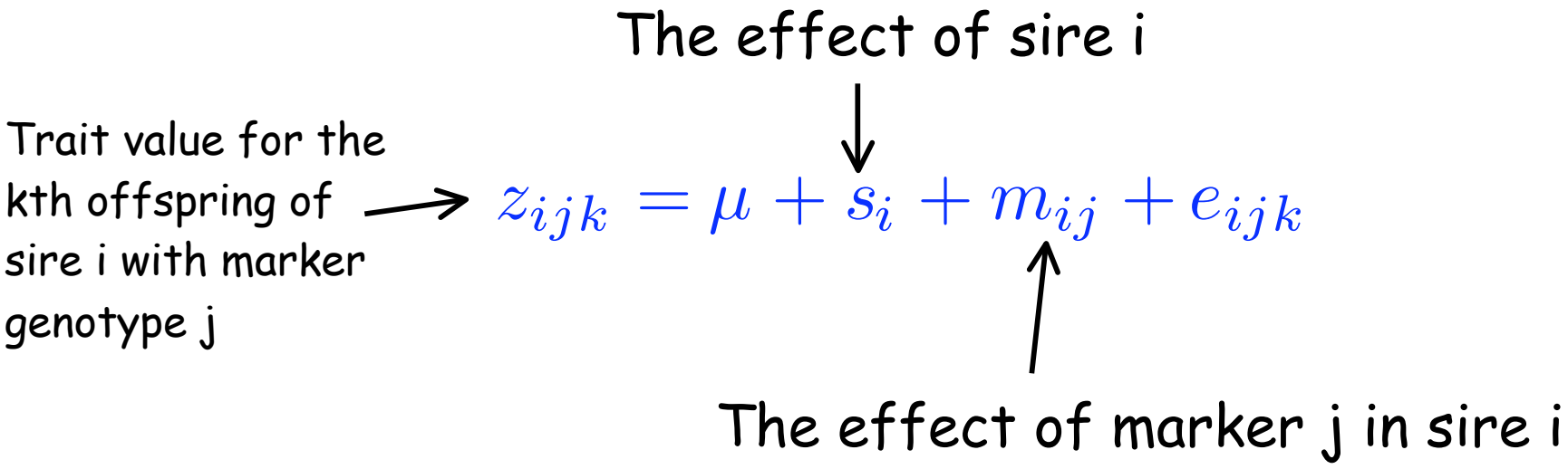
$M_iM_j \times M_iM_j$ Both the same marker heterozygote

Only homozygous offspring informative in distinguishing alternative parental alleles

Sib Families

QTL mapping can occur in sib families. Here, one looks separately within each family for differences in trait means for individuals carrying alternative marker alleles. Hence, a separate analysis can be done for each parent in each family.

Information across families is combined using a standard nested ANOVA. Example: Half-sibs (common sire)



A significant marker effect indicates linkage to a QTL

This is tested using the standard F-ratio, $F = \frac{\mathrm{MS}_m}{\mathrm{MS}_e}$

What can we say about QTL effect and position?

$$\sigma_m^2 = \frac{E(\mathrm{MS}_m) - E(\mathrm{MS}_e)}{n/2} = (1-2c)^2 \frac{\sigma_A^2}{2}$$

Thus, the marker variance confounds both position and QTL effect, here measured by the additive variance of the QTL

$$\sigma_m^2 = \frac{E(\mathrm{MS}_m) - E(\mathrm{MS}_e)}{n/2} = (1-2c)^2 \frac{\sigma_A^2}{2}$$

The marker variance confounds both position and QTL effect, here measured by the additive variance of the QTL

Since $\sigma_A^2 = 2a^2p(1\text{-}p)$, we can get a small variance for a QTL of large effect ($a \gg 1$) if one allele is rare

If $2p(1\text{-}p)$ is small, heterozygotes are rare (most sires are QTL homozygotes). However, if a is large, in these rare families, there is a large effect.

Hence, there is a tradeoff in getting a sufficient number of families to have a few with the QTL segregating, but also to have family sizes large enough to detect differences between parental marker alleles

The granddaughter design. Widely used in dairy cattle to improve power.

Each sire (i) produces a number of sons that are genotyped for the sire allele. Each son then produces a number of offspring in which the trait is measured

Effect of sire i

Effect of the kth son of sire i with sire marker allele j

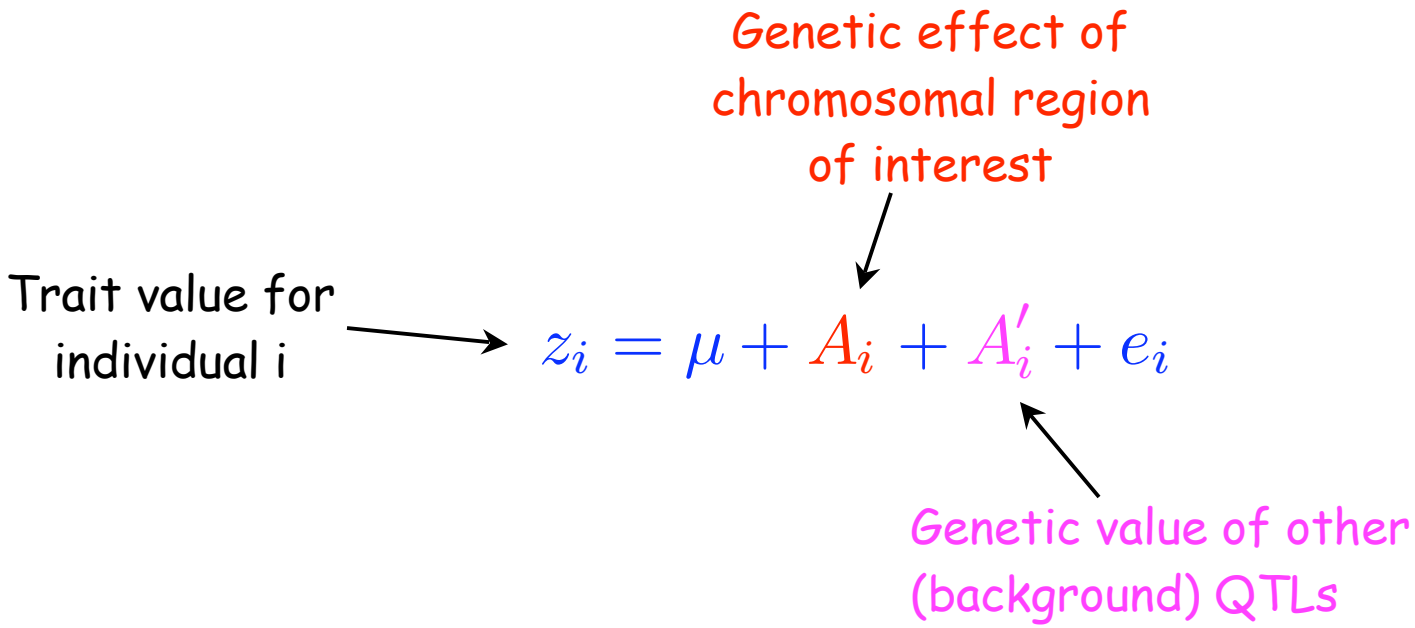
$$z_{ijk\ell} = \mu + g_i + m_{ij} + s_{ijk} + e_{ijk\ell}$$

Effect of marker allele j from sire i

Advantage: large sample size for each m_{ij} value.

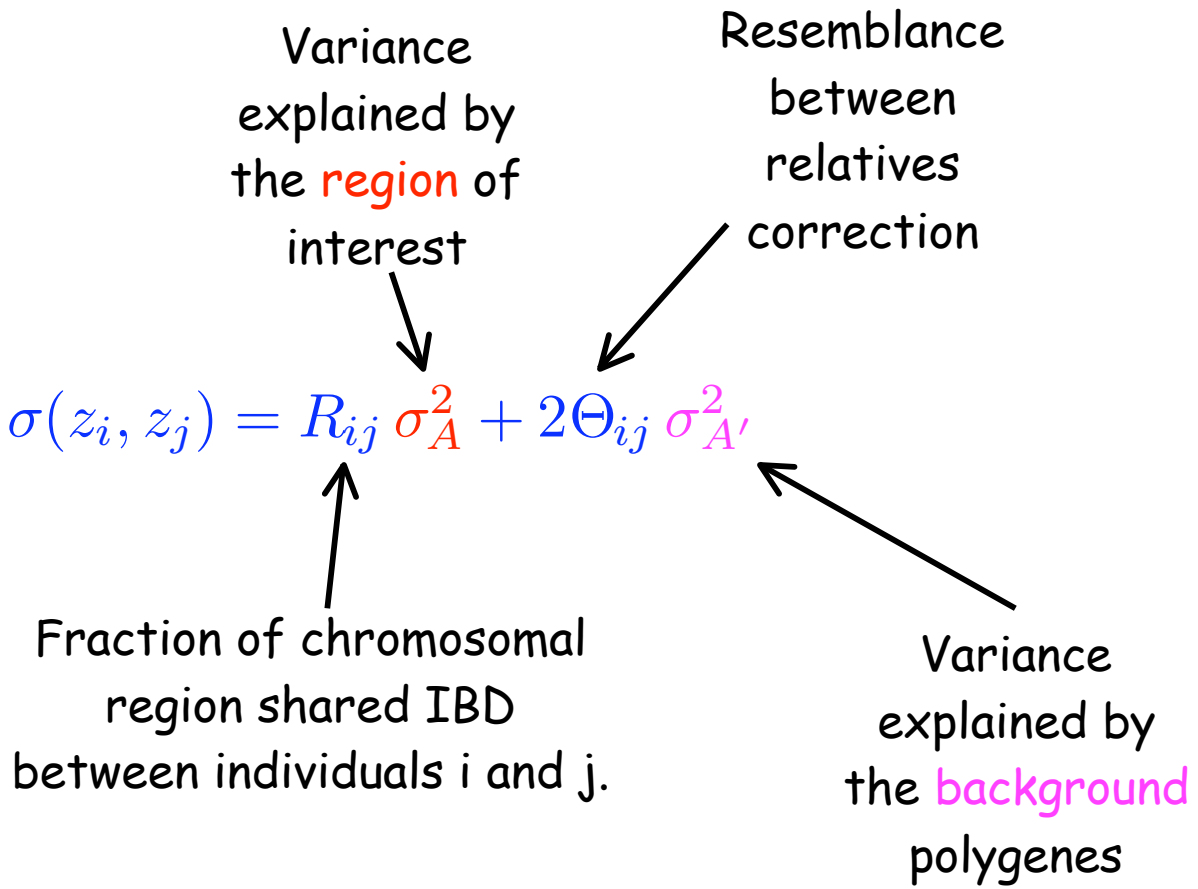
General Pedigree Methods

Random effects (hence, variance component) method for detecting QTLs in general pedigrees



$$z_i = \mu + A_i + A'_i + e_i$$

The covariance between individuals i and j is thus



Assume z is MVN, giving the covariance matrix as

$$\mathbf{V} = \mathbf{R}\,\sigma_A^2 + \mathbf{A}\,\sigma_{A'}^2 + \mathbf{I}\,\sigma_e^2$$

Here

$$\mathbf{R}_{ij} = \begin{cases} 1 & \text{for } i = j \\ \widehat{R}_{ij} & \text{for } i \neq j \end{cases}, \qquad \mathbf{A}_{ij} = \begin{cases} 1 & \text{for } i = j \\ 2\Theta_{ij} & \text{for } i \neq j \end{cases}$$

Estimated from marker data

Estimated from the pedigree

The resulting likelihood function is

$$\ell(\,\mathbf{z}\,|\,\mu, \sigma_A^2, \sigma_{A'}^2, \sigma_e^2\,) = \frac{1}{\sqrt{(2\pi)^n\,|\mathbf{V}|}} \exp\left[-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T\,\mathbf{V}^{-1}\,(\mathbf{z}-\boldsymbol{\mu})\right]$$

A significant σ_A^2 indicates a linked QTL.

Haseman-Elston Regressions

One simple test for linkage of a QTL to a marker locus is the Haseman-Elston regression, used in human genetics

The idea is simple: If a marker is linked to a QTL, then relatives that share a IBD marker alleles likely share IBD QTL alleles and hence are more similar to each other than expected by chance.

The approach: regress the (squared) difference in trait value in the same sets of relatives on the fraction of IBD marker alleles they share.

For the ith pair of relatives,

$$Y_i = (z_{i1} - z_{i2})^2$$

$$Y_i = a + \beta\pi_{im} + e$$

Fraction of marker alleles IBD in this pair of relatives. π = 0, 0.5, or 1

The expected slope β is a function of the additive variance of the linked QTL, the distance c between marker and QTL and the type of relative

$$\beta = \begin{cases} -2(1-2c)\,\sigma_A^2 & \text{grandparent–grandchild;} \\ -2(1-2c)^2\,\sigma_A^2 & \text{half-sibs;} \\ -2(1-2c)^2\,(1-c)\,\sigma_A^2 & \text{avuncular (aunt/uncle–nephew/niece).} \end{cases}$$

$$Y_i = a + \beta \pi_{im} + e$$

$$\beta = \begin{cases} -2\,(1-2c)\,\sigma_A^2 & \text{grandparent–grandchild;} \\ -2\,(1-2c)^2\,\sigma_A^2 & \text{half-sibs;} \\ -2\,(1-2c)^2\,(1-c)\,\sigma_A^2 & \text{avuncular (aunt/uncle–nephew/niece).} \end{cases}$$

This is a one-sided test, as the null hypothesis (no linkage) is β =0 versus the alternative β < 0

Note that parent-offspring are NOT an appropriate pair of relatives for this test. WHY?

Affected Sib Pair Methods

As with the HE regression, the idea is that if the marker is linked to a QTL, individuals with more IBD marker alleles with have closer phenotypes.

Example of an allele-sharing method.

Consider a discrete phenotype (disease presence/ absence).

A sib can either be affected or unaffected.

A pair of sibs can either be concordant (either doubly affected or both unaffected), or discordant (a singly-affected pair)

The IDB probabilities for a random pair of full sibs are Pr(0 IBD) = Pr(2 IBD) = 1/4, Pr(1 IBD) = 1/2

The idea of affected sib pair (ASP) methods is to compare this expected distribution across one (or more) classes of phenotypes. A departure from this expectation implies the marker is linked to a QTL. There are a huge number of versions of this simple test.

p_{ij} = frequency of a pair with i affected sibs sharing j marker alleles IBD

- Compare the frequency of doubly-affected sib pairs that have both marker alleles IBD with the null value 1/4

$$T_2 = \frac{\widehat{p}_{22} - 1/4}{\sqrt{\frac{3}{16\, n_2}}}$$

One-sided test as $p_{22} > 1/4$ under linkage

- Compare the mean number of IBD alleles in doubly-affected pairs with the null value of 1

One-sided test as $p_{21} + 2p_{22} > 1$ under linkage

Number of doubly-affected pairs

$$T_m = \sqrt{2\, n_2\, (\,\widehat{p}_{21} + 2\,\widehat{p}_{22} - 1\,)}$$

Freq of pairs sharing 1 allele IBD

Freq of pairs with 2 alleles IBD

Finally, maximum likelihood approaches have been suggested. In particular, the goodness-of-fit of the full distribution of IDB values in doubly-affected sibs

Comparing $p_{2i} = n_{2i}/n_2$ with 1/4 (i=0, 2) or 1/2 (i=1)

The resulting test statistic is called MLS for Maximum LOD score, and is given by

$$MLS = \log_{10}\left[\prod_{i=0}^{2}\left(\frac{\widehat{p}_{2i}}{\pi_{2i}}\right)^{n_{2i}}\right] = \sum_{i=0}^{2} n_{2i}\,\log_{10}\left(\frac{\widehat{p}_{2i}}{\pi_{2i}}\right)$$

Linkage indicated by MLS > 3

An alternative formulation of MLS is to consider just the contribution from one parent, where (under no linkage), Pr(sibs share same parental allele IBD) = Pr(sibs don't share same parental allele) = 1/2

$$MLS = (1 - n_1)\log_{10}\left(\frac{1-\widehat{p_1}}{1/2}\right) + n_1 \log_{10}\left(\frac{\widehat{p_1}}{1/2}\right)$$

Fraction of sib pairs sharing parental allele IBD

Example: Genomic scan for type I diabetes

Marker D6S415. For sibs with diabetes, 74 pairs shared the same parental allele IBD, 60 did not

$$MLS[\text{D6S415}] = 60\ \log_{10}\left[\frac{2\ \text{x}\ 60}{134}\right] + 74\ \log_{10}\left[\frac{2\ \text{x}\ 74}{134}\right] = 0.32$$

Marker D6S273. For sibs with diabetes, 92 pairs shared the same parental allele IBD, 31 did not

$$MLS[\text{D6S273}] = 31\ \log_{10}\left[\frac{2\ \text{x}\ 31}{123}\right] + 92\ \log_{10}\left[\frac{2\ \text{x}\ 92}{123}\right] = 6.87$$

Association & LD mapping

Mapping major genes (LD mapping) vs. trying to Map QTLs (Association mapping)

Idea: Collect random sample of individuals, contrast Trait means over marker genotypes

If a dense enough marker map, likely population level linkage disequilibrium (LD) between closely-linked genes

LD: Linkage disequilibrium

$D(AB) = freq(AB) - freq(A)*freq(B)$.

LD = 0 if A and B are independent. If LD not zero, correlation between A and B in the population

If a marker and QTL are linked, then the marker and QTL alleles are in LD in close relatives, generating a marker-trait association.

The decay of D: $D(t) = (1-c)^t D(0)$

here c is the recombination rate. Tightly-linked genes (small c) initially in LD can retain LD for long periods of time

Linkage disequilibrium mapping

Idea is to use a random sample of individuals from the population rather than a large pedigree.

Ironically, in the right settings this approach has more power for fine mapping than pedigree analysis.

Why?

Key is the expected number of recombinants. in a pedigree, Prob(no recombinants) in n individuals is $(1-c)^n$

LD mapping uses the historical recombinants in a sample. Prob(no recomb) = $(1-c)^{2t}$, where t = Time back to most recent common ancestor

Expected number of recombinants in a sample of n sibs is cn

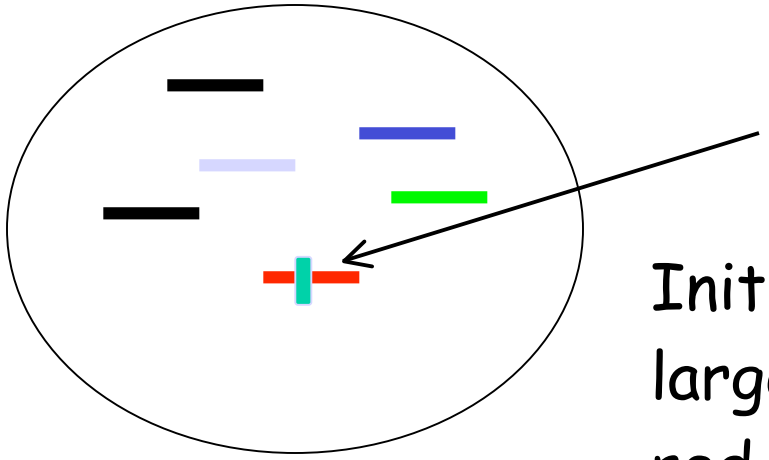
Expected number of recombinants in a sample of n random individuals with a time t back to the MRCA (most recent common ancestor) is $2cnt$

Hence, if t is large, many more expected recombinants in random sample and hence more power for very fine mapping (i.e. $c < 0.01$)

Because so many expected recombinants, only works with c very small

Fine-mapping genes

Suppose an allele causing a large effect on the trait arose as a single mutation in a closed population



New mutation arises on red chromosome

Initially, the new mutation is largely associated with the red haplotype

Hence, markers that define the red haplotype are likely to be associated (i.e. in LD) with the mutant allele

This linkage disequilibrium decays slowly with time if c is small

Let π = Prob(mutation associated with original haplotype)

$$\pi = (1-c)^t$$

Thus if we can estimate π and t, we can solve for c,

$$c = 1 - \pi^{1/t}$$

Diastrophic dysplasis (DTD) association with CSF1R marker locus alleles

Allele	Normal	DTD-bearing
1-1	4 (3.3%)	144 (94.7%)
1-2	28 (22.7%)	1 (0.7%)
2-1	7 (5.7%)	0 (0%)
2-2	84 (68.3%)	7 (4.6%)

Most frequent allele type varies between normal and DTD-bearing haplotypes

Hence, allele 1-1 appears to be on the original haplotype in which the DTD mutation arose --> $\pi = 0.947$

Diastrophic dysplasis (DTD) association with CSF1R marker locus alleles

Allele	Normal	DTD-bearing
1-1	4 (3.3%)	144 (94.7%)
1-2	28 (22.7%)	1 (0.7%)
2-1	7 (5.7%)	0 (0%)
2-2	84 (68.3%)	7 (4.6%)

$\pi = 0.947$

$$c = 1 - \pi^{1/t} = 1 - 0.947^{1/100}$$

100 generations to MRCA used for Finnish population

Gives c = 0.00051 between marker and DTD. Best estimate from pedigrees is c = 0.012 (1.2cM)

Candidate Loci and the TDT

Often try to map genes by using case/control contrasts, also called association mapping.

The frequencies of marker alleles are measured in both a

case sample -- showing the trait (or extreme values)

control sample -- not showing the trait

The idea is that if the marker is in tight linkage, we might expect LD between it and the particular DNA site causing the trait variation.

Problem with case-control approach: Population Stratification can given false positives.

When population being sampled actually consists of several distinct subpopulations we have lumped together, marker alleles may provide information as to which group an individual belongs. If there are other risk factors in a group, this can create a false association btw marker and trait

Example. The Gm marker was thought (for biological reasons) to be an excellent candidate gene for diabetes in the high-risk population of Pima Indians in the American Southwest. Initially a very strong association was observed:

Gm^+	Total	% with diabetes
Present	293	8%
Absent	4,627	29%

Gm^+	Total	% with diabetes
Present	293	8%
Absent	4,627	29%

Problem: freq(Gm^+) in Caucasians (lower-risk diabetes Population) is 67%, Gm^+ rare in full-blooded Pima

The association was re-examined in a population of Pima that were 7/8th (or more) full heritage:

Gm^+	Total	% with diabetes
Present	17	59%
Absent	1,764	60%

Transmission-disequilibrium test (TDT)

The TDT accounts for population structure. It requires sets of relatives and compares the number of times a marker allele is transmitted (T) versus not-transmitted (NT) from a marker heterozygote parent to affected offspring.

Under the hypothesis of no linkage, these values should be equal, resulting in a chi-square test for lack of fit:

$$\chi^2_{td} = \frac{(T - NT)^2}{(T + NT)}$$

Scan for type I diabetes in Humans. Marker locus D2S152

Allele	T	NT	χ^2	p
228	81	45	10.29	0.001
230	59	73	1.48	0.223
240	36	24	2.30	0.121

$$\chi^2 = \frac{(81 - 45)^2}{(81 + 45)} = 10.29$$

Linkage vs. Association

The distinction between linkage and association is subtle, yet critical

Marker allele M is associated with the trait if

$Cov(M,y) \neq 0$

While such associations can arise via linkage, they can also arise via population structure.

Thus, association DOES NOT imply linkage, and linkage is not sufficient for association

Linkage within each family generates a non-random distribution of transmitted alleles

Population-level disequilibrium is also required for the TDT as we average over alleles. If linkage phase significantly varies over families, no effect

How much such population-level disequilibrium be generated?

If the disease is caused by a single mutation, it started out in a particular genetic background (haplotype), and hence in linkage disequilibrium.

Over time, this decays, but for very tightly linked markers the time to decay can be considerable.

Dense SNP Association Mapping

Mapping genes using known sets of relatives can be problematic because of the cost and difficulty in obtaining enough relatives to have sufficient power.

By contrast, it is straightforward to gather large sets of unrelated individuals, for example a large number of cases (individuals with a particular trait/disease) and controls (those without it).

With the very dense set of SNP markers (dense = very tightly linked), it is possible to scan for markers in LD in a random mating population with QTLs, simply because c is so small that LD has not yet decayed

These ideas lead to consideration of a strategy of

.

For example, using 30,000 equally spaced SNP in The 3000cM human genome places any QTL within 0.05cM of a SNP. Hence, for an association created t generations ago (for example, by a new mutant allele appearing at that QTL, the fraction of original LD still present is at least $(1-0.0005)^t \sim 1-\exp(t*0.0005)$. Thus for mutations 100, 500, and 1000 generations old (2.5K, 12.5K, and 25 K years for humans), this fraction is 95.1%, 77.8%, 60.6%,

We thus have large samples and high disequilibrium, the recipe needed to detect linked QTLs of small effect

Given that we can easily increase the density of SNPs (even over our 30K example), and that increasing the number of individuals sampled is usually not a serious problem, we are still faced with the problem of population structure

Two general classes of approaches have been proposed to deal with with.

(i) Attempts to correct for the common population structure signal (Genomic Control, regression approaches)

(ii) Attempts to first assign individuals into subpopulations, with association tests then conducted within each (Structured Association Mapping)

Genomic Control

Devlin and Roeder (1999). Basic idea is that association tests (marker presence/absence vs. trait presence/absence) is typically done with a standard 2 x 2 χ^2 test.

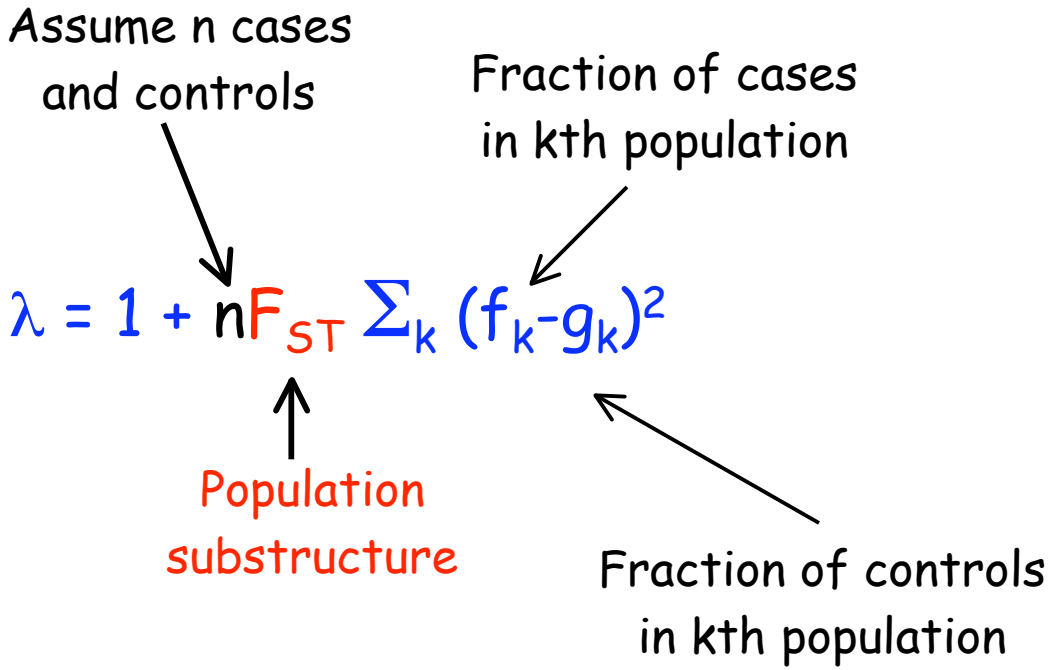
When population structure is present, the test statistic now follows a scaled χ^2, so that if S is the test statistic, then $S/\lambda \sim \chi^2_1$ (so $S \sim \lambda\chi^2_1$)

The inflation factor λ is given by

$$\lambda = 1 + nF_{ST} \Sigma_k (f_k - g_k)^2$$

Note that this departure from a χ^2 increases with sample size n

Genomic Control



Genomic control attempts to estimate λ directly from our distribution of test statistics S

Estimation of λ

The mean of a χ^2_1 is one. Hence, since S ~ $\lambda\chi^2_1$ and we expect most test statistic values to be from the null (no linkage), one estimator of λ is simply the mean of S, the mean value of the test statistics.

The problem is that this is not a particular robust estimator, as a few extreme values of S (as would occur with linkage!) can inflate λ over its true value.

A more robust estimator is offered from the medium (50% value) of the test statistics, so that for m tests

$$\widehat{\lambda} = \frac{\text{medium}\,(S_1, \cdots; S_m)}{0.456}$$

Structured Association Mapping

Pritchard and Rosenberg (1999) proposed Structured Association Mapping, wherein one assumes k subpopulations (each in Hardy-Weinberg).

Given a large number of markers, one then attempts to assign individuals to groups using an MCMC Bayesian classifier

Once individuals assigned to groups, association mapping without any correction can occur in each group.

Structure plus Kinship Methods

Association mapping in plants offer occurs by first taking a large collection of lines, some closely related, others more distantly related. Thus, in addition to this collection being a series of subpopulations (derivatives from a number of founding lines), there can also be additional structure within each subpopulation (groups of more closely related lines within any particular lineage).

$$Y = X\beta + Sa + Qv + Zu + e$$

Fixed effects in blue, random effects in red

Q-K method

$$Y = X\beta + Sa + Qv + Zu + e$$

β = vector of fixed effects

a = SNP effects

v = vector of subpopulation effects (STRUCTURE)

Q_{ij} = Prob(individual I in group j). Determined from STRUCTURE output

u = shared polygenic effects due to kinship.

Cov(u) = var(A)*A, where the relationship matrix A estimated from marker data matrix K

Regression Approaches

A third approach to control for structure is simply to include a number of markers, outside of the SNP of interest, chosen because they are expected to vary over any subpopulations

How might you choose these in a sample? Try those markers (read STRs) that show the largest departure from Hardy-Weinberg, as this is expected in markers that vary the most over subpopulations.

Indicator (0 / 1) Variable
for SNP genotype k. Typically
k = 3, i.e. AA, Aa aa

$$y = \mu + \sum_{k=1}^{n} \beta_k M_k + \sum_{j=1}^{m} \gamma_j b_j + e$$

Significant β indicates
marker-trait association

SNP marker
under consideration

m unlinked markers that
vary across subpopulations.
b_j = marker genotype indicator
variable