Lecture 2:
QTL and Association mapping

Bruce Walsh
Second Bangalore School of
Population Genetics and Evolution
25 Jan- 5 Feb 2016

Part I
QTL mapping and the use of inbred line crosses

- QTL mapping tries to detect small (20-40 cM) chromosome segments influencing trait variation
  - Relatively crude level of resolution
- QTL mapping performed either using inbred line crosses or sets of known relatives
  - Uses the simple fact of an excess of parental gametes

Inbred lines

$$\frac{M\ Q}{M\ Q} \times \frac{m\ q}{m\ q} \rightarrow F_1\ \frac{M\ Q}{m\ q} \rightarrow$$

gametes	freq
M Q	(1-c)/2
m q	(1-c)/2
M q	c/2
m Q	c/2

Inbred lines

$$\frac{M\ Q}{M\ Q} \times \frac{m\ q}{m\ q} \rightarrow F_1\ \frac{M\ Q}{m\ q} \rightarrow$$

gametes	freq
M Q	0.49
m q	0.49
M q	0.01
m Q	0.01

c = 0.02

Creates a marker-trait association in offspring, with M-bearing chromosomes co-segregating with Q, so that M-bearing gametes will (on average) yield larger trait values (here 98% of M are Q)

Key idea: Looking for marker-trait associations in collections of relatives

If (say) the mean trait value for marker genotype MM is statistically different from that for genotype mm, then the M/m marker is linked to a QTL

One can use a random collection of such markers spanning a genome (a genomic scan) to search for QTLs

Conditional Probabilities of QTL Genotypes

The basic building block for all QTL methods is $Pr(Q_k \mid M_j)$ --- the probability of QTL genotype Q_k given the marker genotype is M_j.

$$Pr(Q_k \mid M_j) = \frac{Pr(Q_k M_j)}{Pr(M_j)}$$

Consider a QTL linked to a marker (recombination Fraction = c). Cross MMQQ x mmqq. In the F1, all gametes are MQ and mq

In the F2, freq(MQ) = freq(mq) = (1-c)/2,
freq(mQ) = freq(Mq) = c/2

Hence, Pr(MMQQ) = Pr(MQ)Pr(MQ) = $(1-c)^2/4$

Pr(MMQq) = 2Pr(MQ)Pr(Mq) = $2c(1-c)/4$

Pr(MMqq) = Pr(Mq)Pr(Mq) = $c^2/4$

Why the 2? MQ from father, Mq from mother, OR MQ from mother, Mq from father

Since Pr(MM) = 1/4, the conditional probabilities become

Pr(QQ | MM) = Pr(MMQQ)/Pr(MM) = $(1-c)^2$

Pr(Qq | MM) = Pr(MMQq)/Pr(MM) = $2c(1-c)$

Pr(qq | MM) = Pr(MMqq)/Pr(MM) = c^2

How do we use these?

Expected Marker Means

The expected trait mean for marker genotype M_j is just

$$\mu_{M_j} = \sum_{k=1}^{N} \mu_{Q_k} \Pr(Q_k \mid M_j)$$

For example, if QQ = 2a, Qq = a(1+k), qq = 0, then in the F2 of an MMQQ/mmqq cross,

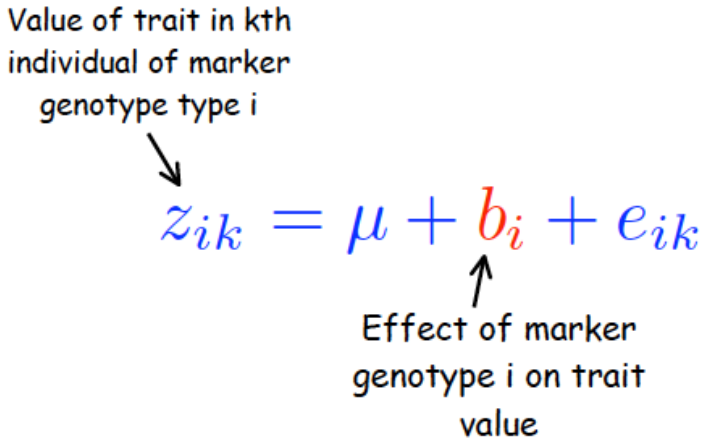
$$(\mu_{MM} - \mu_{mm})/2 = a(1-2c)$$

- If the trait mean is significantly different for the genotypes at a marker locus, it is linked to a QTL
- A small MM-mm difference could be (i) a tightly-linked QTL of small effect or (ii) loose linkage to a large QTL

Linear Models for QTL Detection

The use of differences in the mean trait value for different marker genotypes to detect a QTL and estimate its effects is a use of linear models.

One-way ANOVA.



$$z_{ik} = \mu + b_i + e_{ik}$$

Detection: a QTL is linked to the marker if at least one of the b_i is significantly different from zero

Estimation: (QTL effect and position): This requires relating the b_i to the QTL effects and map position

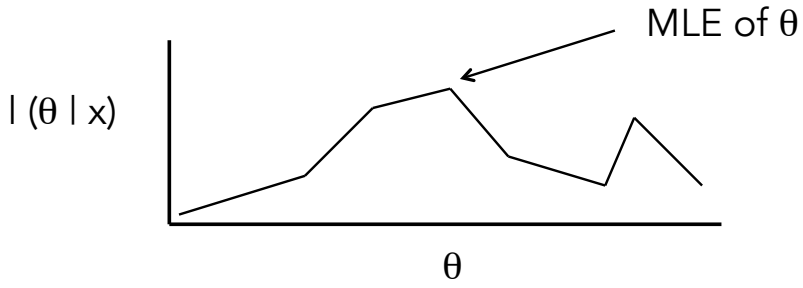
Maximum Likelihood

$p(x_1,\ldots, x_n \mid \theta)$ = density of the observed data $(x_1,\ldots, x_n)$ given the (unknown) distribution parameter(s) θ

Fisher suggested the method of maximum likelihood given the data $(x_1,\ldots, x_n)$ find the value(s) of θ that maximize $p(x_1,\ldots, x_n \mid \theta)$

We usually express $p(x_1,\ldots, x_n \mid \theta)$ as a likelihood function $l(\theta \mid x_1,\ldots, x_n)$ to remind us that it is dependent on the observed data

The Maximum Likelihood Estimator (MLE) of θ are the value(s) that maximize the likelihood function l given the observed data $x_1,\ldots, x_n$.



This is formalized by looking at the log-likelihood surface, $L = \ln [l(\theta \mid x)]$. Since ln is a monotonic function, the value of θ that maximizes l also maximizes L

The curvature of the likelihood surface in the neighborhood of the MLE informs us as to the precision of the estimator. A narrow peak = high precision. A board peak = low precision

$$\text{Var(MLE)} = -1 \Big/ \frac{\partial^2 L(\mu \mid \mathbf{z})}{\partial \mu^2}$$

The larger the curvature, the smaller the variance

Likelihood Ratio tests

Hypothesis testing in the ML frameworks occurs through likelihood-ratio (LR) tests

$$LR = 2\ln\left(\frac{\ell(\hat{\Theta}_r \mid \mathbf{z})}{\ell(\hat{\Theta} \mid \mathbf{z})}\right) = 2\left[L(\hat{\Theta}_r \mid \mathbf{z}) - L(\hat{\Theta} \mid \mathbf{z})\right]$$

θ_r is the MLE under the restricted conditions (some parameters specified, e.g., var =1)

Θ is the MLE under the unrestricted conditions (no parameters specified)

For large sample sizes (generally) LR approaches a Chi-square distribution with r df (r = number of parameters assigned fixed values under null)

Maximum Likelihood Methods

Sum over the N possible linked QTL genotypes

Probability of QTL genotype k given marker genotype j --- genetic map and linkage phase enter here

$$\ell(z \mid M_j) = \sum_{k=1}^{N} \varphi(z, \mu_{Q_k}, \sigma^2) \Pr(Q_k \mid M_j)$$

Distribution of trait value given QTL genotype is k is normal with mean μ_{Qk}. (QTL effects enter here)

ML methods combine both detection and estimation of QTL effects/position.

Test for a linked QTL given from by the Likelihood Ratio (or LR) test

Maximum of the likelihood under a no-linked QTL model

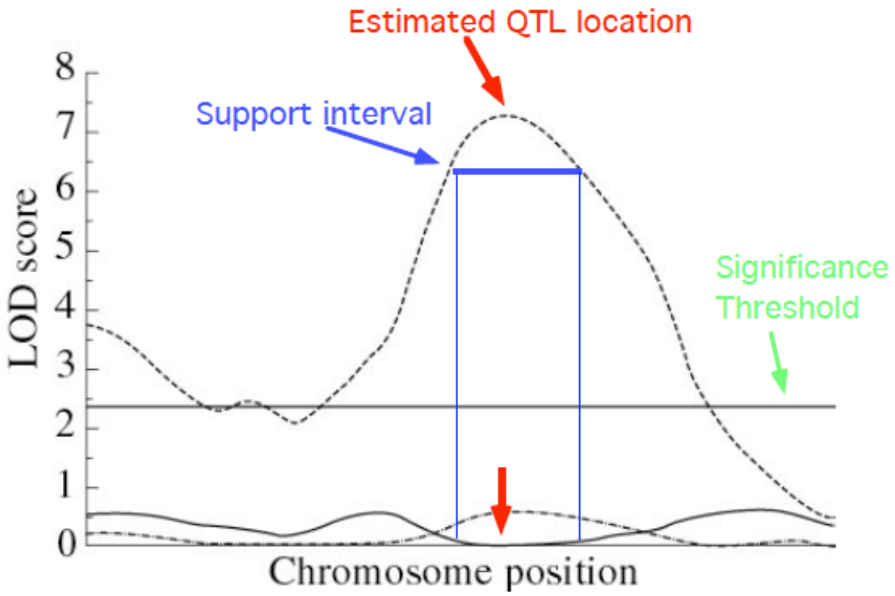
$$\mathrm{LR} = -2 \ln \frac{\max \ell_r(\mathbf{z})}{\max \ell(\mathbf{z})}$$

Maximum of the full likelihood

The LR score is often plotted by trying different locations for the QTL (i.e., values of c) and computing a LOD score for each

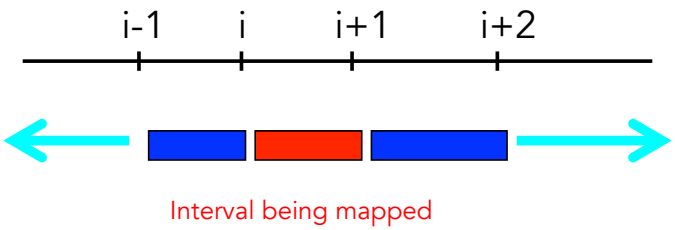
$$\mathrm{LOD}(c) = -\log_{10}\left[\frac{\max \ell_r(\mathbf{z})}{\max \ell(\mathbf{z}, c)}\right] = \frac{\mathrm{LR}(c)}{2 \ln 10} \simeq \frac{\mathrm{LR}(c)}{4.61}$$

A typical QTL map from a likelihood analysis

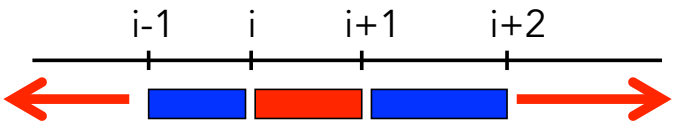


Interval Mapping with Marker Cofactors

Consider interval mapping using the markers i and i+1. QTLs linked to these markers, but outside this interval, can contribute (falsely) to estimation of QTL position and effect



Now suppose we also add the two markers flanking the interval (i-1 and i+2)



Inclusion of markers i-1 and i+2 fully account for any linked QTLs to the left of i-1 and the right of i+2

Interval mapping + marker cofactors is called Composite Interval Mapping (CIM)

CIM works by adding an additional term to the linear model,

$$\sum_{k \neq i, i+1} b_k \, x_{kj}$$

CIM also (potentially) includes unlinked markers to account for QTL on other chromosomes.

Power and Precision

While modest sample sizes are sufficient to detect a QTL of modest effect (power), large sample sizes are required to map it with any precision

With 200-300 F_2, a QTL accounting for 5% of total variation can be mapped to a 40cM interval

Over 10,000 F_2 individuals are required to map this QTL to a 1cM interval

Power and Repeatability: The Beavis Effect

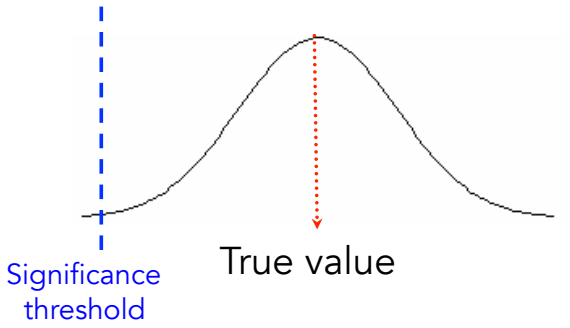
QTLs with low power of detection tend to have their effects ***overestimated***, often very dramatically

As power of detection increases, the overestimation of detected QTLs becomes far less serious

This is often called the Beavis Effect, after Bill Beavis who first noticed this in simulation studies. This phenomena is also called the winner's curse in statistics (and GWAS)

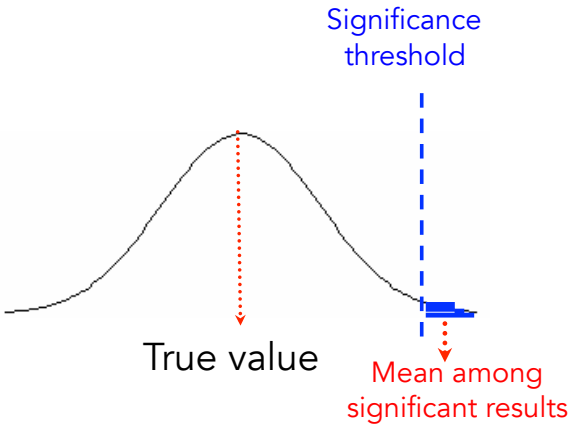
Beavis Effect

Also called the "winner's curse" in the GWAS literature

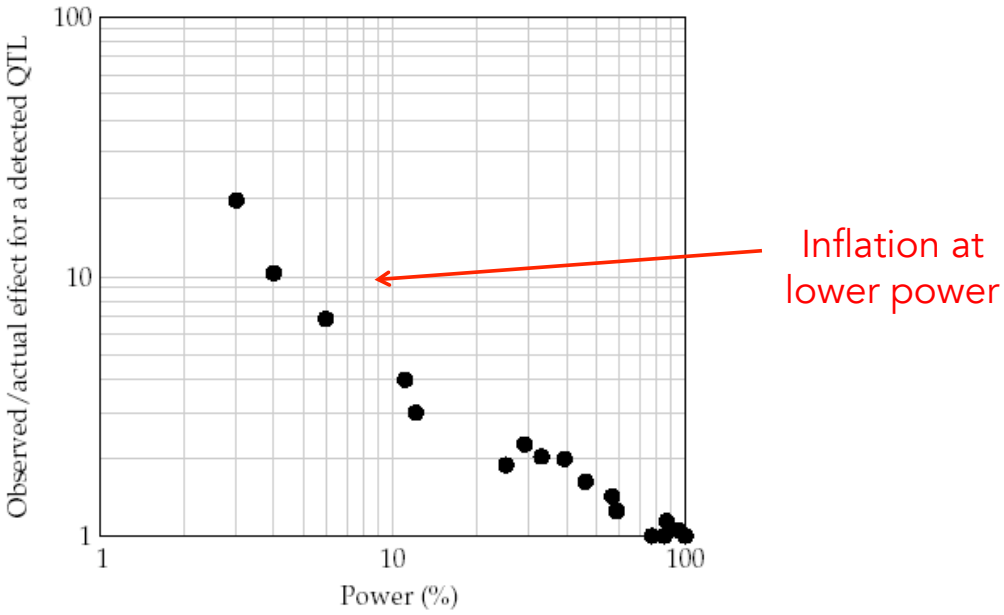


High power setting: Most realizations are to the right of the significance threshold. Hence, the average value given the estimate is declared significant (above the threshold) is very close to the true value.

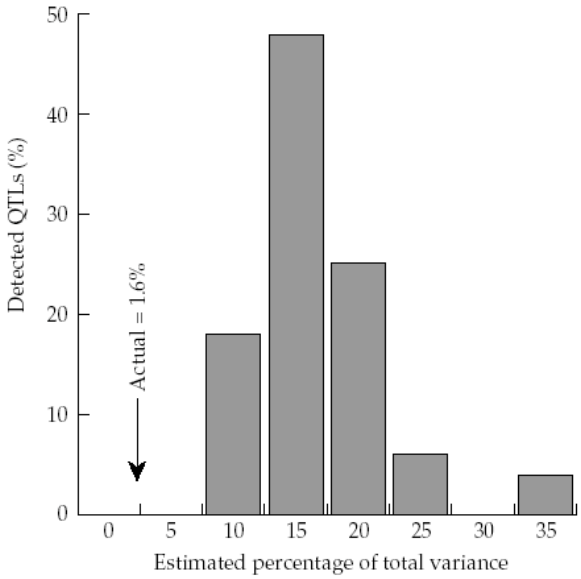
In low power settings, most realizations are below the significance threshold, hence most of the time the effect is scored as being nonsignificant



However, the mean of those declared significant is much larger than the true mean



Inflation can be significant, esp. with low power



Beavis simulation: actual effect size is 1.6% of variation. Estimated effects (at significant markers) much higher

What is a "QTL"

- A detected "QTL" in a mapping experiment is a region of a chromosome detected by linkage.
- Usually large (typically 10-40 cM)
- When further examined, most "large" QTLs turn out to be a linked collection of locations with increasingly smaller effects
- The more one localizes, the more subregions that are found, and the smaller the effect in each subregion
- This is called fractionation

Limitations of QTL mapping

- Poor resolution (~20 cM or greater in most designs with sample sizes in low to mid 100's)
  - Detected "QTLs" are thus large chromosomal regions
- Fine mapping requires either
  - Further crosses (recombinations) involving regions of interest (i.e., RILs, NILs)
  - Enormous sample sizes
    - If marker-QTL distance is 0.5cM, require sample sizes in excess of 3400 to have a 95% chance of 10 (or more) recombination events in sample
    - 10 recombination events allows one to separate effects that differ by ~ 0.6 SD

Limitations of QTL mapping (cont)

- "Major" QTLs typically fractionate
  - QTLs of large effect (accounting for > 10% of the variance) are routinely discovered.
  - However, a large QTL peak in an initial experiment generally becomes a series of smaller and smaller peaks upon subsequent fine-mapping.
- The Beavis effect:
  - When power for detection is low, marker-trait associations declared to be statistically significant significantly overestimate their true effects.
  - This effect can be very large (order of magnitude) when power is low.

II: QTL mapping in Outbred Populations and Association Mapping

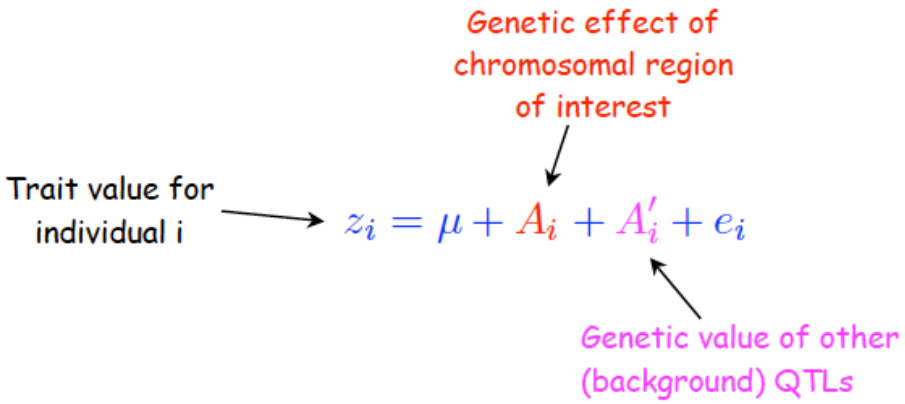
- Association mapping uses a set of very dense markers in a set of (largely) unrelated individuals
- Requires population level LD
- Allows for very fine mapping (1-20 kB)

QTL mapping in outbred populations

- Much lower power than line-cross QTL mapping
- The gametes from each parent must be separately analyzed for marker-trait associations
- We focus on an approach for general pedigrees, as this leads us into association mapping

General Pedigree Methods

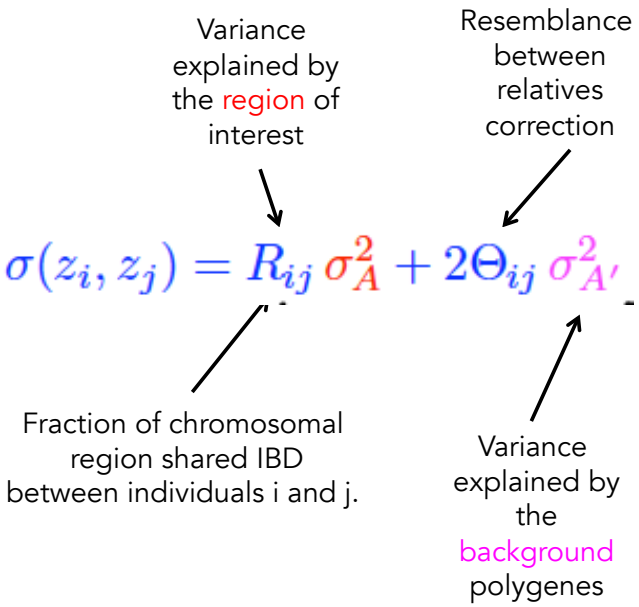
Random effects (hence, variance component) method for detecting QTLs in general pedigrees



The model is rerun for each marker

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

The covariance between individuals i and j is thus



Assume z is MVN, giving the covariance matrix as

$$\mathbf{V} = \mathbf{R}\,\sigma^2_A + \mathbf{A}\,\sigma^2_{A'} + \mathbf{I}\,\sigma^2_e$$

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

$$\ell(\mathbf{z}\,|\,\mu, \sigma^2_A, \sigma^2_{A'}, \sigma^2_e) = \frac{1}{\sqrt{(2\pi)^n\,|\mathbf{V}|}} \exp\left[-\frac{1}{2}(\mathbf{z}-\mu)^T\,\mathbf{V}^{-1}\,(\mathbf{z}-\mu)\right]$$

A significant σ_A^2 indicates a linked QTL.

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

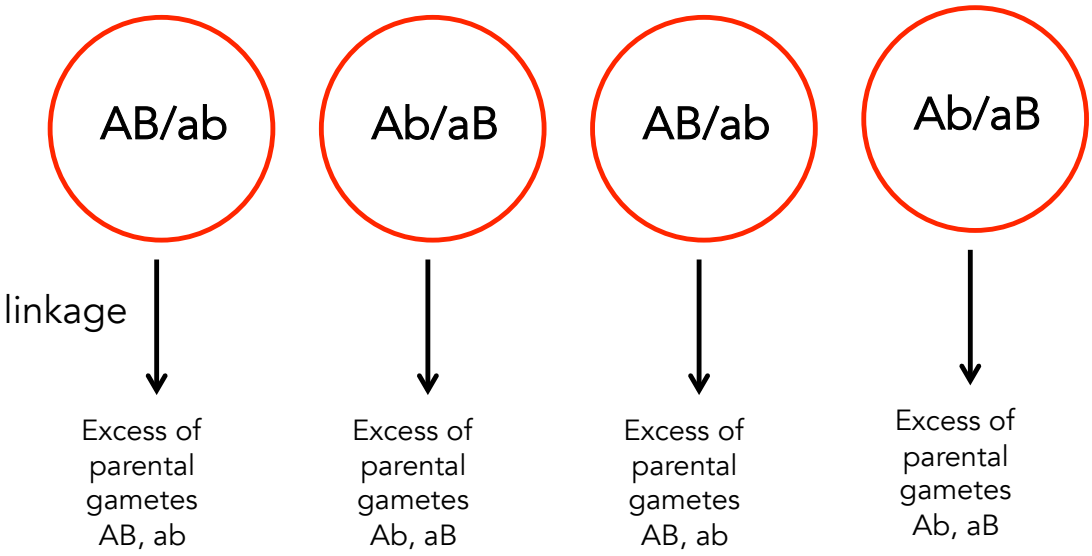
LD: Linkage disequilibrium

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.

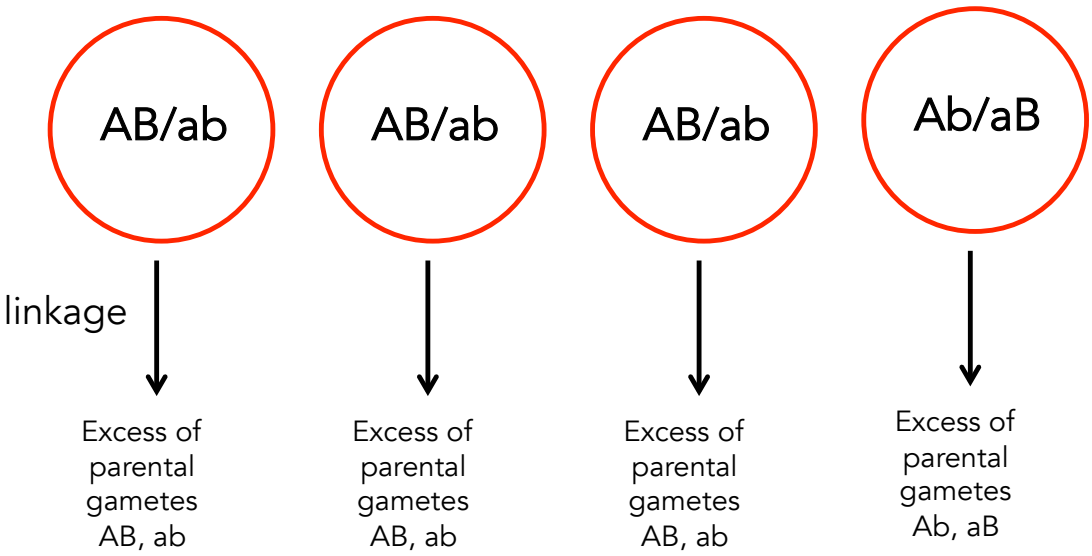
The decay of D: $D(t) = (1-c)^t D(0)$
here c is the recombination rate. Tightly-linked genes (small c) initially in LD can retain LD for long periods of time

No LD: random distribution of linkage phases



Pool all gametes: AB, ab, Ab, aB equally frequent

With LD, nonrandom distribution of linkage phase



Pool all gametes: Excess of AB, ab due to an excess of AB/ab parents

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%,

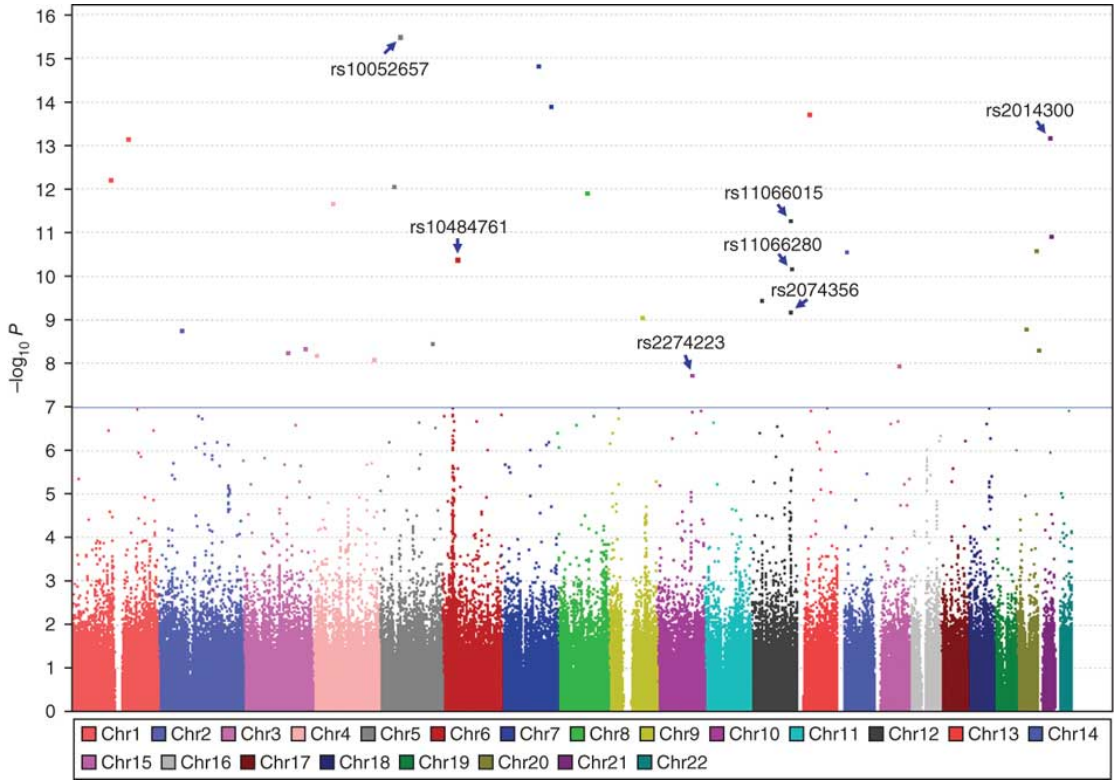
We thus have large samples and high disequilibrium, the recipe needed to detect linked QTLs of small effect

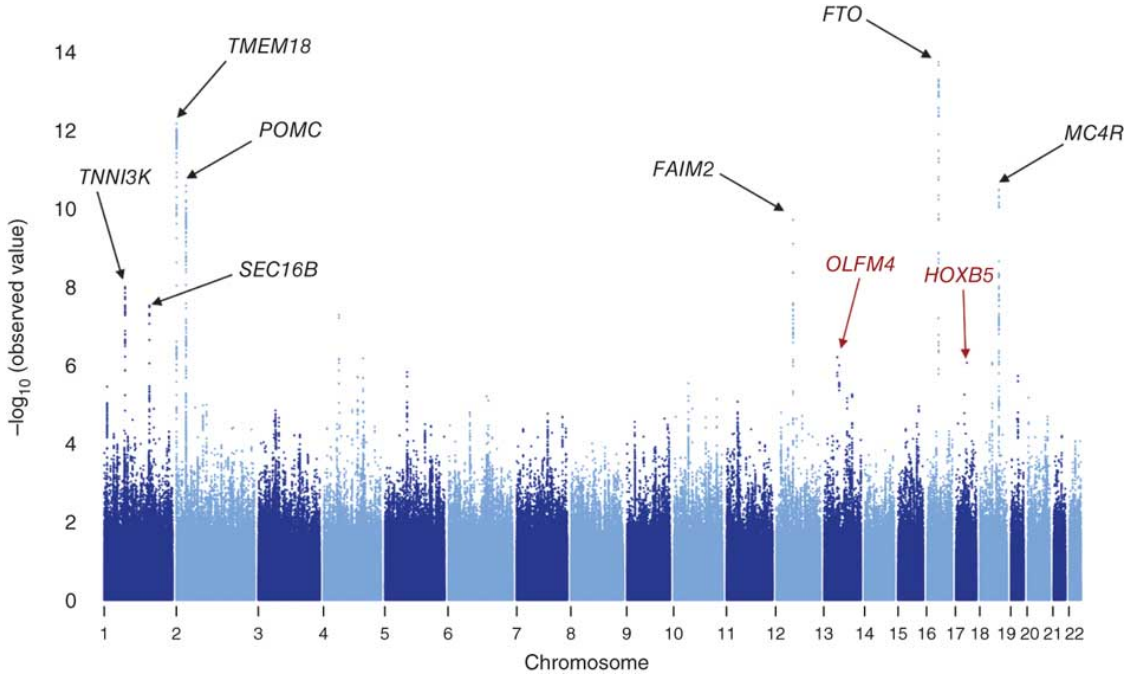
Association mapping

- Marker-trait associations within a population of unrelated individuals
- Very high marker density (~ 100s of markers/cM) required
  - Marker density no less than the average track length of linkage disequilibrium (LD)
- Relies on very slow breakdown of initial LD generated by a new mutation near a marker to generate marker-trait associations
  - LD decays very quickly unless very tight linkage
  - Hence, resolution on the scale of LD in the population(s) being studied (1 ~ 40 kB)
- Widely used since mid 1990's. Mainstay of human genetics, strong inroads in breeding, evolutionary genetics
- Power a function of the genetic variance of a QTL, not its mean effects

Manhattan plots

- The results for a Genome-wide Association study (or GWAS) are typically displayed using a Manhattan plot.
  - At each SNP, -ln(p), the negative log of the p value for a significant marker-trait association is plotted. Values above a threshold indicate significant effects
  - Threshold set by Bonferroni-style multiple comparisons correction
  - With n markers, an overall false-positive rate of p requires each marker be tested using p/n.
  - With n = 10^6 SNPs, p must exceed $0.01/10^6$ or 10^{-8} to have a control of 1% of a false-positive


rs10052657
rs2014300
rs11066015
rs10484761
rs11066280
rs2074356
rs2274223
−log10 P
Chr1 Chr2 Chr3 Chr4 Chr5 Chr6 Chr7 Chr8 Chr9 Chr10 Chr11 Chr12 Chr13 Chr14
Chr15 Chr16 Chr17 Chr18 Chr19 Chr20 Chr21 Chr22


FTO
TMEM18
POMC
MC4R
TNNI3K
FAIM2
SEC16B
OLFM4
HOXB5
−log10 (observed value)
Chromosome

Candidate Loci

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 (and association mapping in general): Population Stratification can give 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%

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) is not 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

Accounting for population structure

- Three classes of approaches proposed
  - 1) Attempts to correct for common pop structure signal (genomic control, regression/ PC methods)
  - 2) Attempts to first assign individuals into subpopulations and then perform association mapping in each set (Structure)
  - 3) Mixed models that use all of the marker information (Tassle, EMMA, many others)
    - These can also account for cryptic relatedness in the data set, which also causes false-positives.

Structured Association Mapping

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

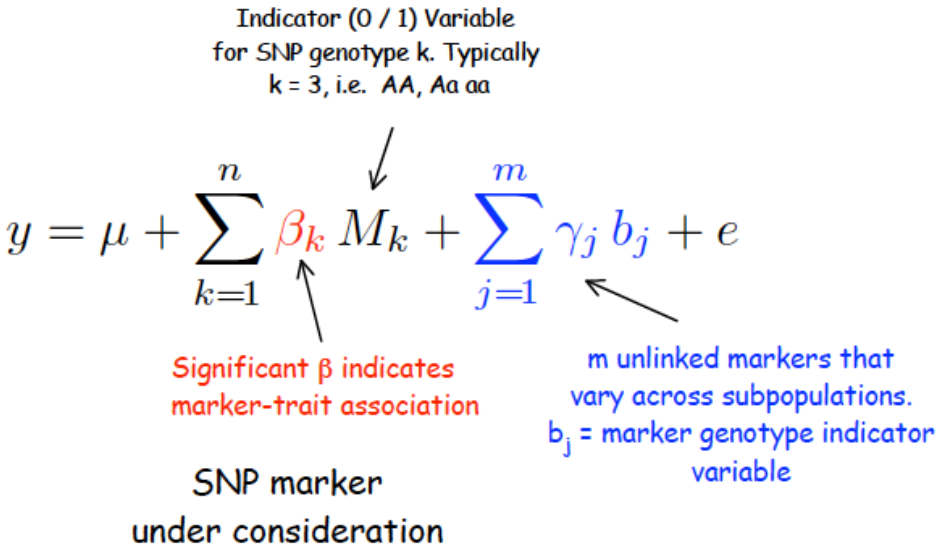
Given a large number of markers, one then attempts to assign individuals to groups using an MCMC Bayesian classifier

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

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.



Variations on this theme (eigenstrat) --- use all of the marker information to extract a set of significant PCs, which are then included in the model as cofactors

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

This is a mixed-model approach. The program TASSEL runs this model.

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, also called a GRM – a genomic relationship matrix

Which markers to include in K?

- Best approach is to leave out the marker being tested (and any in LD with it) when construction the genomic relationship matrix
  - LOCO approach – leave out one chromosome (which the tested marker is linked to)
- Best approach seems to be to use most of the markers
- Other mixed-model approaches along these lines

Power of Association mapping

Q/q is the polymorphic site contributing to trait variation, M/m alleles (at a SNP) used as a marker

Let p be the frequency of M, and assume that Q only resides on the M background (complete disequilibrium)

Haloptype	Frequency	effect
QM	rp	a
qM	(1-r)p	0
qm	1-p	0

Haloptype	Frequency	effect
QM	rp	a
qM	(1-r)p	0
qm	1-p	0

Effect of m = 0

Effect of M = ar

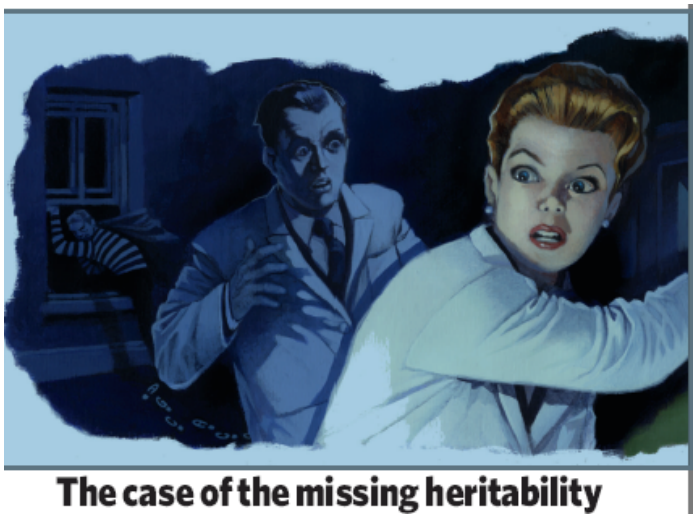
Genetic variation associated with Q = $2(rp)(1-rp)a^2$ ~ $2rpa^2$ when Q rare. Hence, little power if Q rare

Genetic variation associated with marker M is $2p(1-p)(ar)^2 \sim 2pa^2r^2$

Ratio of marker/true effect variance is ~ r

Hence, if Q rare within the A class, even less power!

"How wonderful that we have met with a paradox. Now we have some hope of making progress" -- Neils Bohr



The "missing heritability" pseudo-paradox

- A number of GWAS workers noted that the sum of their significant marker variances was much less (typically 10%) than the additive variance estimated from biometrical methods
- The "missing heritability" problem was birthed from this observation.
- Not a paradox at all
  - Low power means small effect (i.e. variance) sites are unlikely to be called as significant, esp. given the high stringency associated with control of false positives over tens of thousands of tests
  - Further, even if all markers are detected, only a fraction ~ r (the frequency of the causative site within a marker haplotype class) of the underlying variance is accounted for.

GBLUP

- The Q-K method tests SNPs one at a time, treating them as fixed effects
- The general pedigree method (slides 35-36) also tests one marker at a time, treating them as random effects
- Genomic selection can be thought of as estimating all of the SNP effects at once and hence can also be used for GWAS

BLUP, GBLUP, and GWAS

- Pedigree information gives EXPECTED value of shared sites (i.e., ½ for full-sibs)
  - A matrix in BLUP
  - The actual realization of the fraction of shared genes for a particular pair of relatives can be rather different, due to sampling variance in segregation of alleles
  - GRM, genomic relationship matrix (or K or marker matrix M)
  - Hence "identical" relatives can differ significantly in faction of shared regions
  - Dense marker information can account for this

The general setting

- Suppose we have n measured individuals (the n x 1 vector **y** of trait values)
- The n x n relationship matrix **A** gives the relatedness among the sampled individuals, where the elements of **A** are obtained from the pedigree of measured individuals
- We may also have p (>> n) SNPs per individual, where the n x p marker information matrix **M** contains the marker data, where M_{ij} = score for SNP j (i.e., 0 for 00, 1 for 10, 2 for 11) in individual i.

Covariance structure of random effects

- A critical element specifying the mixed model is the covariance structure (matrix) of the vector **u** of random effects
- Standard form is that Cov(**u**) = variance component * matrix of known constants
  - This is the case for pedigree data, where **u** is typically the vector of breeding values, and the pedigree defines a relationship matrix **A**, with Cov(**u**) = Var(A) * **A**, the additive variance times the relationship matrix
  - With marker data, the covariance of random effects are functions of the marker information matrix **M**.
    - If **u** is the vector of p marker effects, then Cov(**u**) = Var(m) * $\mathbf{M}^T\mathbf{M}$, the marker variance times the covariance structure of the markers.

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

Pedigree-based BV estimation: (BLUP)
u_{nx1} = vector of BVs, Cov(u) = Var(A) A_{nxn}

Marker-based BV estimation: (GBLUP)
u_{nx1} = vector of BVs, Cov(u) = Var(m) M^TM (n x n)

GWAS: u_{px1} = vector of marker effects,
Cov(u) = Var(m) MM^T (p x p)

Genomic selection: predicted vector of breeding values from marker effects (genetic breeding values),
$GBV_{nx1} = M_{nxp}u_{px1}$.
Note that Cov(GBV) = Var(m) M^TM (n x n)

Many variations of these general ideas by adding additional assumptions on covariance structure.

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

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

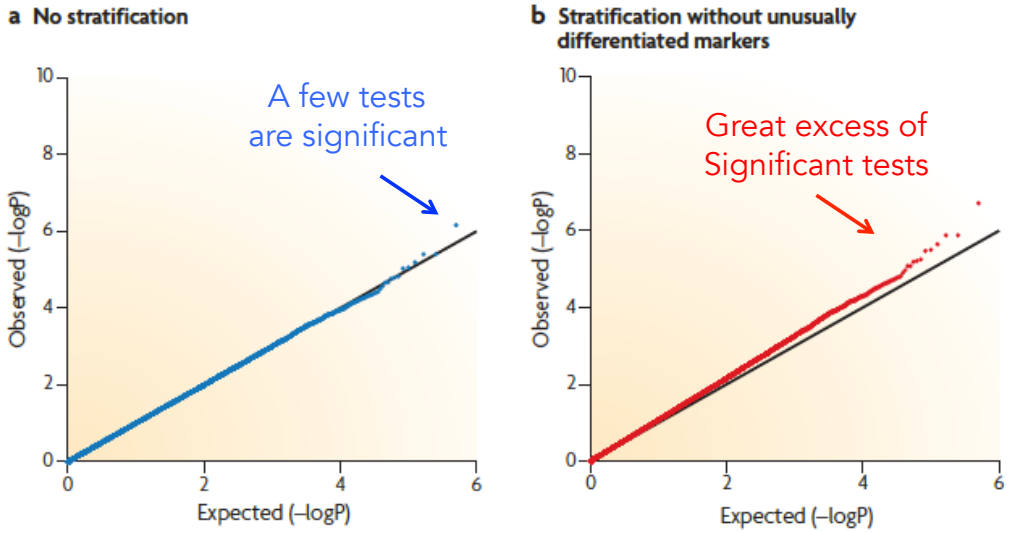
GWAS Model diagnostics

Genomic control λ as a diagnostic tool

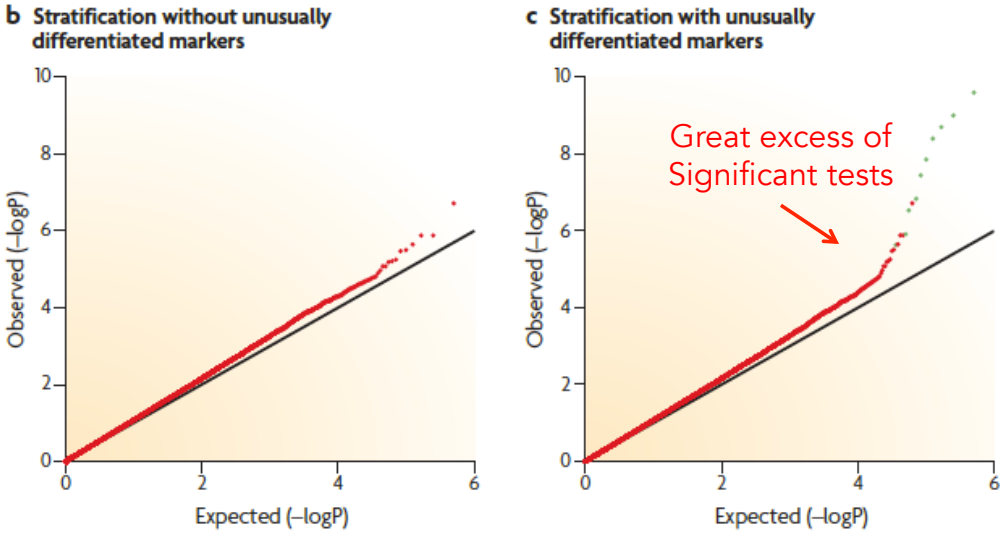
- Presence of population structure will inflate the λ parameter
- A value above 1 is considered evidence of additional structure in the data
  - Could be population structure, cryptic relatedness, or both
  - A lambda value less that 1.05 is generally considered benign
- One issue is that if the true polygenic model holds (lots of sites of small effect), then a significant fraction will have inflated p values, and hence an inflated λ value.
- Hence, often one computes the λ following attempts to remove population structure. If the resulting value is below 1.05, suggestion that structure has been largely removed.

P – P plots

- Another powerful diagnostic tool is the p-p plot.
- If all tests are drawn from the null, then the distribution of p values should be uniform.
  - There should be a slight excess of tests with very low p indicating true positives
- This gives a straight line of a log-log plot of observed (seen) and expected (uniform) p values with a slight rise near small values
  - If the fraction of true positives is high (i.e., many sites influence the trait), this also bends the p-p plot



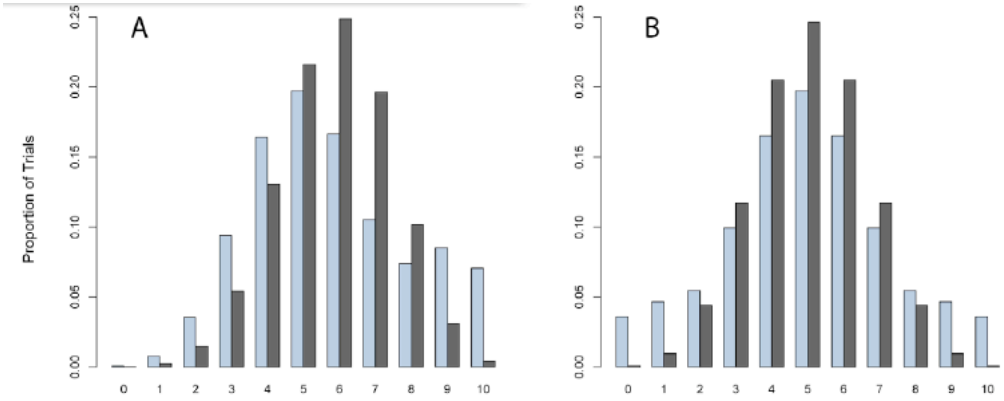
Price et al. 2010 Nat Rev Gene 11: 459



As with using λ, one should construct p-p following some approach to correct for structure & relatedness to see if they look unusual.

Dealing with Rare Variants

- Many disease may be influenced by rare variants.
  - Problem: Each is rare and thus overall gives a weak signal, so testing each variant is out (huge multiple-testing problem)
  - However, whole-genome sequencing (or just sequencing through a target gene/region) is designed to pick up such variants
- Burden tests are one approach
  - Idea: When comparing case vs. controls, is there an overdispersion of mutations between the two categories?



Solid = random distribution over cases/controls
Blue = observed distribution

A: Variants only increase disease risk (excess at high values)

B: Variants can both increase (excess high values) and decrease risk (excess low values) --- inflation of the variance σ^2

C(α) test

- Idea: Suppose a fraction p_0 of the sample are controls, $p_1 = 1-p_0$ are cases. Note these varies are fixed over all variants
- Let n_i be the total number of copies of a rare variant i.
- Under binomial sampling, the expected number of variant i in the case group is ~ Bin(p_1,n_i)
- Pool the observations of all such variants over a gene/region of interest and ask if the variance in the number in cases exceeds the binomial sampling variance $n_i p_1(1-p_1)$

C(α) test (cont).

- Suppose m variants in a region, test statistic is of the form
- $\Sigma_i (y_i - n_i p_1)^2 - n_i p_1 (1-p_1)$
- y_i = number of variant I in cases.
- This is observed variance minus binomial prediction
- This is scaled by a variance term to give a test statistic that is roughly normally distributed