

Lecture 7

QTL and Association Mapping with Mixed Models

Bruce Walsh lecture notes
Introduction to Mixed Models
SISG (Module 12), Seattle
17 – 19 July 2019

QTL & Association mapping

- We would like to know both the genomic locations (map positions) and effects (either genotypic means or variances) for genes underlying quantitative trait variation
- QTL mapping
 - Using linkage information on a set of known relatives
- Association mapping
 - Using very fine scale LD to map genes in a set of random individuals from a population

Outline

- Basics of QTL mapping
 - Line crosses
 - typically fixed effects models
 - Outbred populations
 - Random effects family models
 - General pedigree methods
- High parameter models
 - Shrinkage approaches for detecting epistasis
- Association mapping

Inbred Line Cross QTL mapping

- Most powerful design
 - Cross two fully inbred lines, look at marker-trait segregation in the F_2 (or other, such as F_n) generations
 - P1: MMQQ, P2:mmqq
 - All F_1 same genotype/phase: MQ/mq
 - Hence, in the F_1 , all parents have the same genotype
 - At most only two alleles, each with freq 1/2
 - Idea: Does the mean trait value of (say) MM individuals differ from (say) mm
 - Different marker genotypes have different mean trait values

Expected Marker Means

The expected trait mean for marker genotype M_j is just

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

For example, if $QQ = 2a$, $Qq = a(1+d)$, $qq = 0$, 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.

Value of trait in kth
individual of marker
genotype type i



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



Effect of marker
genotype i on trait
value

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

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

Detecting epistasis


One major advantage of linear models is their flexibility. To test for epistasis between two QTLs, use ANOVA with an interaction term

$$z = \mu + a_i + b_k + d_{ik} + e$$

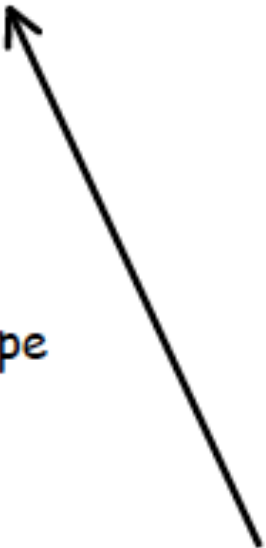
Effect from marker genotype
at first marker set (can be > 1 loci)



Effect from marker genotype
at second marker set



Interaction between marker genotypes i in 1st
marker set and k in 2nd marker set



Detecting epistasis

$$z = \mu + a_i + b_k + d_{ik} + e$$

- At least one of the a_i significantly different from 0
---- QTL linked to first marker set
- At least one of the b_k significantly different from 0
---- QTL linked to second marker set
- At least one of the d_{ik} significantly different from 0
---- interactions between QTL in sets 1 and two

Problem: Huge number of potential interaction terms (order m^2 , where m = number of markers)

Model selection

- With (say) 300 markers, we have (potentially) 300 single-marker terms and $300 \times 299 / 2 = 44,850$ epistatic terms
 - Hence, a model with up to $p = 45,150$ possible parameters
 - 2^p possible submodels = $10^{13,600}$ ouch!
- The issue of **Model selection** becomes very important.
- How do we find the best model?
 - Stepwise regression approaches
 - Forward selection (add terms one at a time)
 - Backwards selection (delete terms one at a time)
 - Try all models, assess best fit
 - Mixed-model approaches (Stochastic Search Variable Selection, or SSVS)

Model Selection

Model Selection: Use some criteria to choose among a number of candidate models. Weight goodness-of-fit (L , value of the likelihood at the MLEs) vs. number of estimated parameters (k)

AIC = Akaike's information criterion

$$AIC = 2k - 2 \ln(L)$$

BIC = Bayesian information criterion (Schwarz criterion)

$$BIC = k \cdot \ln(n)/n - 2 \ln(L)/n$$

BIC penalizes free parameters more strongly than AIC

Other measures. For these (and AIVC, BIC) smaller score indicates better model fit

Model averaging

Model averaging: Generate a composite model by weighting (averaging) the various models, using AIC, BIC, or other

Idea: Perhaps no “best” model, but several models all extremely close. Better to report this “distribution” rather than the best one

One approach is to average the coefficients on the “best-fitting” models using some scheme to return a composite model

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**

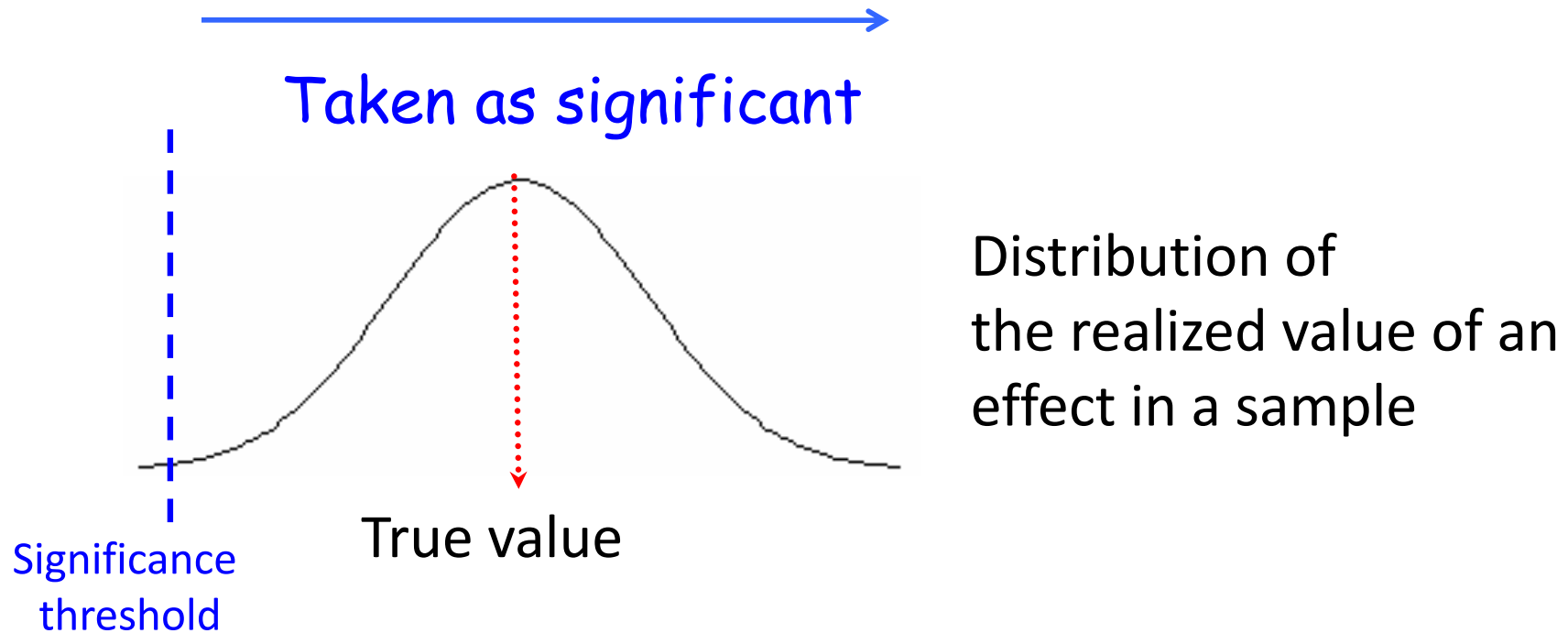
Limitations of QTL mapping

- **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)

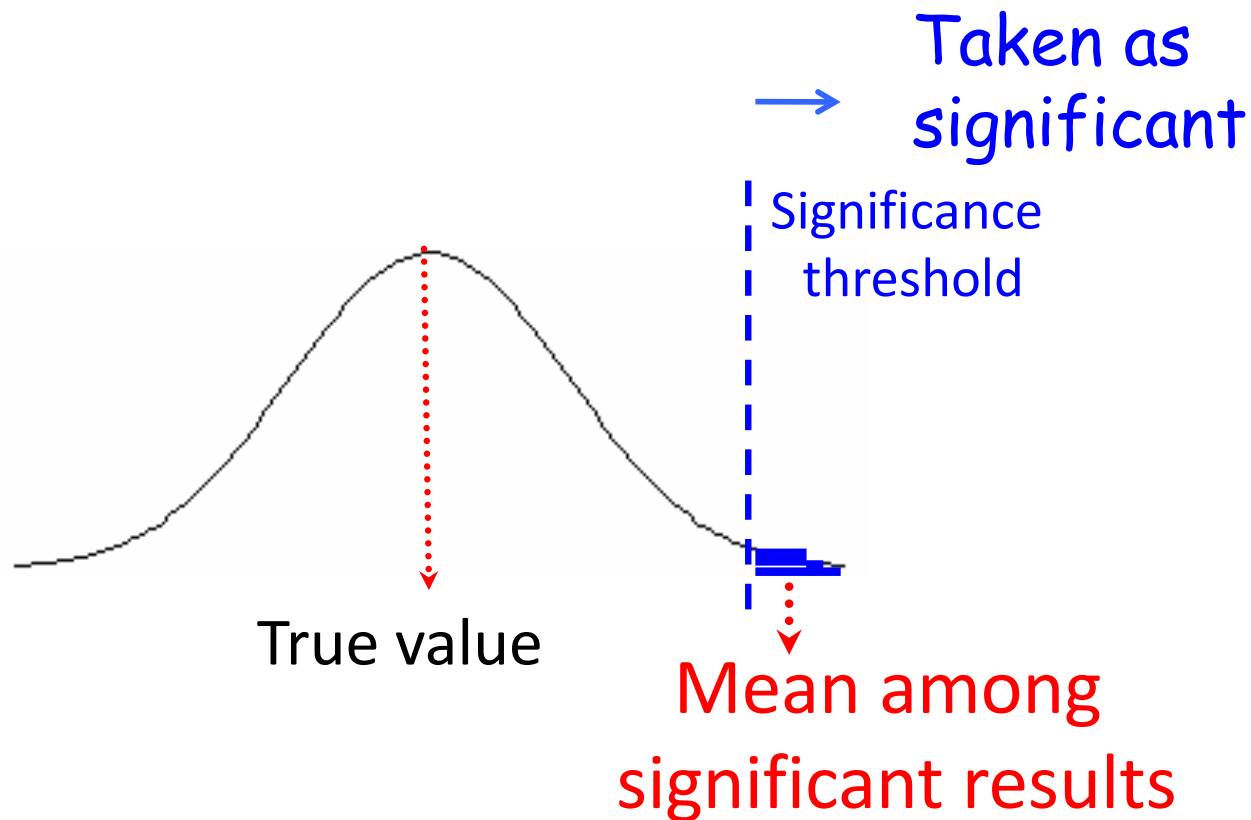
- “Major” QTLs typically **fractionate**
 - QTLs of large effect (accounting for $> 10\%$ of the variance) are routinely discovered.
 - However, a large QTL peak in an initial experiment generally becomes a series of smaller and smaller peaks upon subsequent fine-mapping.
- The **Beavis effect**:
 - 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.

Beavis Effect



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.

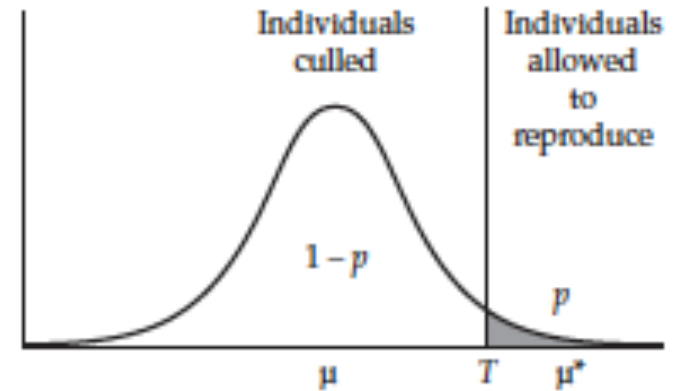
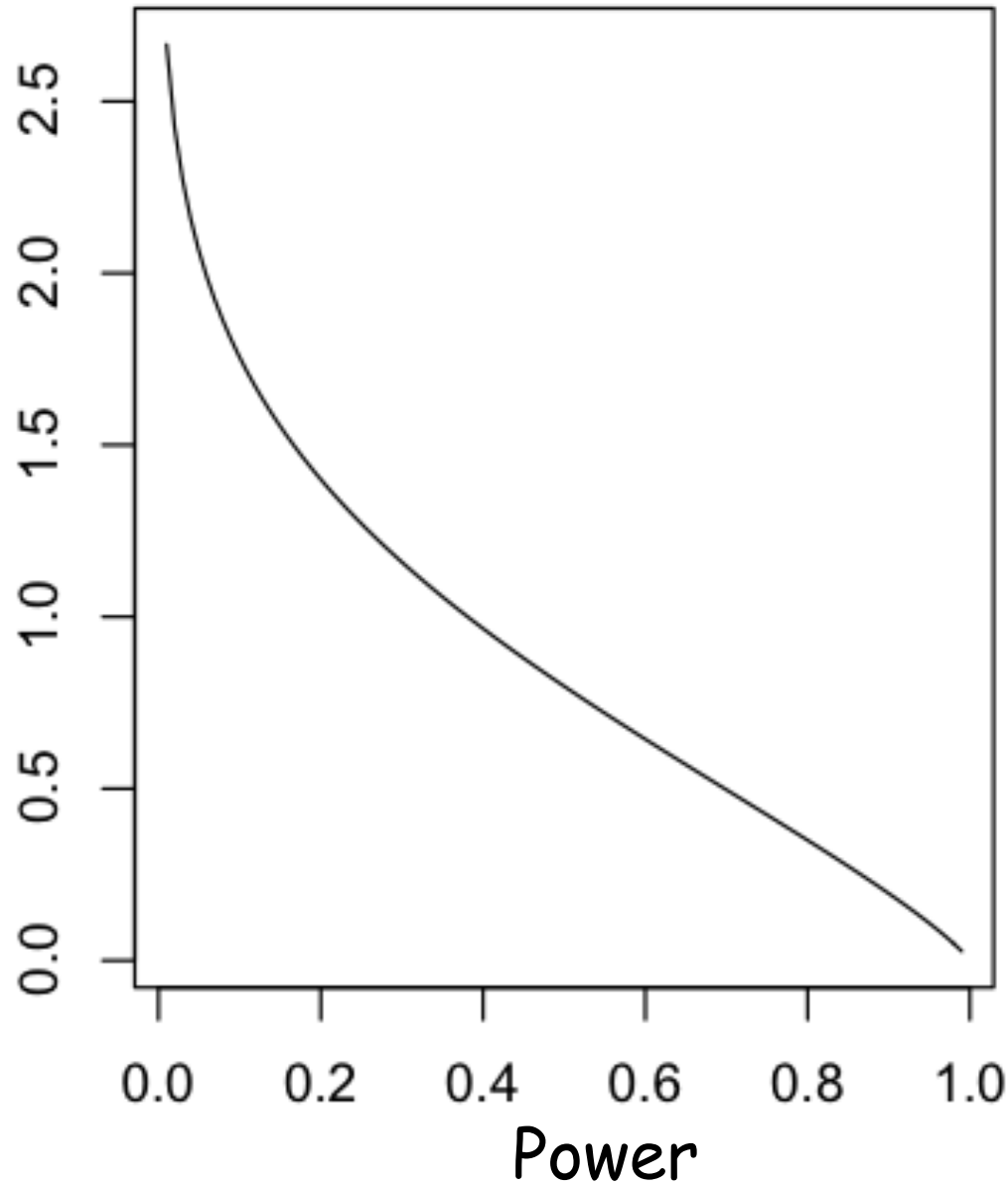
In **low power settings**, most realizations are below the significance threshold, hence most of the time the effect is scored as being nonsignificant



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

Beavis effect is akin to a selection intensity

Overestimation (in SDs)



$$\bar{i} = \frac{S}{\sigma} = \frac{\varphi(x_{[1-p]})}{p}$$

Outbred populations

- When we move from the simple framework of an inbred line cross QTL design to a set of parents from an outbred population, complications arise as the parents don't all have the same genotypes
 - Differences in linkage phase
 - Many uninformative as to linkage (varies over makers)
 - Possibility of multiple alleles
- Result: express marker effects in terms of the variance in trait value it explains, rather than in terms of mean marker effects

General Pedigree Methods

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

Trait value for individual i → $z_i = \mu + A_i + A'_i + e_i$

Genetic effect of chromosomal region of interest

Genetic value of other (background) QTLs

The diagram shows the equation $z_i = \mu + A_i + A'_i + e_i$ with arrows pointing to each term. The text 'Trait value for individual i' points to z_i . The text 'Genetic effect of chromosomal region of interest' points to A_i . The text 'Genetic value of other (background) QTLs' points to A'_i . The text 'Genetic effect of chromosomal region of interest' is in red, and 'Genetic value of other (background) QTLs' is in purple.

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

Variance explained by the **region** of interest

Resemblance between relatives correction

$$\sigma(z_i, z_j) = R_{ij} \sigma_A^2 + 2\Theta_{ij} \sigma_{A'}^2$$

Fraction of chromosomal region shared IBD between individuals i and j.

Variance explained by the **background** polygenes

Assume \mathbf{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

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

Estimated from marker
data

Estimated from
the pedigree

The resulting likelihood function is

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

A significant σ_A^2 indicates a linked QTL.

Association & LD mapping

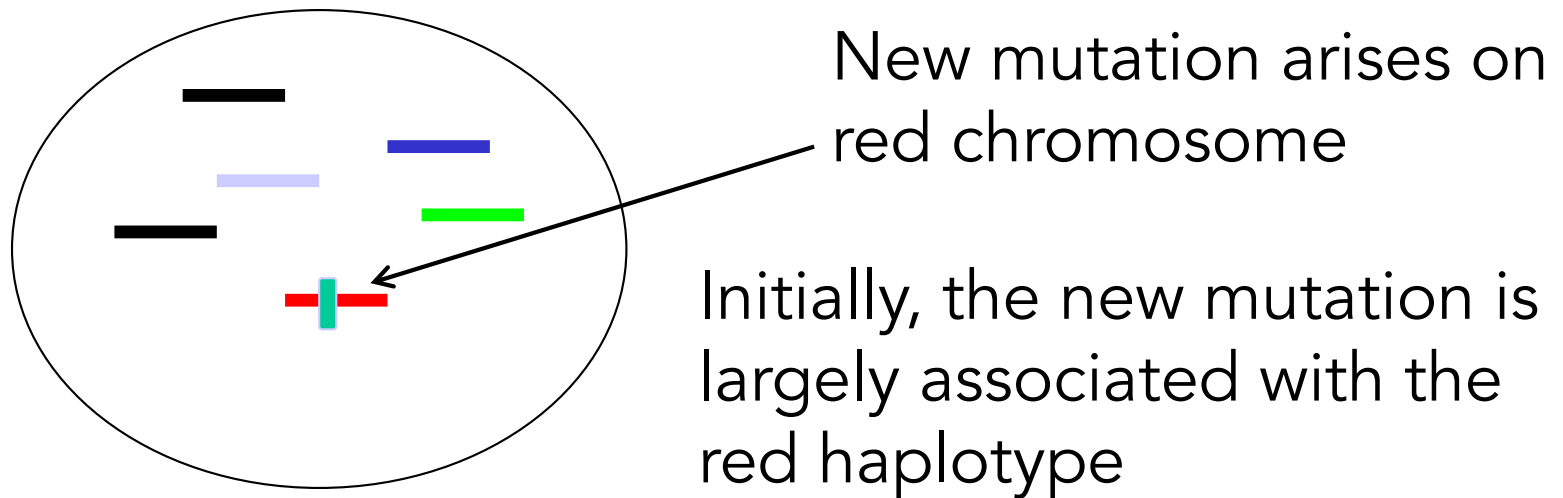
Mapping major genes (LD mapping) vs. trying to Map QTLs (Association mapping)

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

Fine-mapping genes

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



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

Background: Association mapping

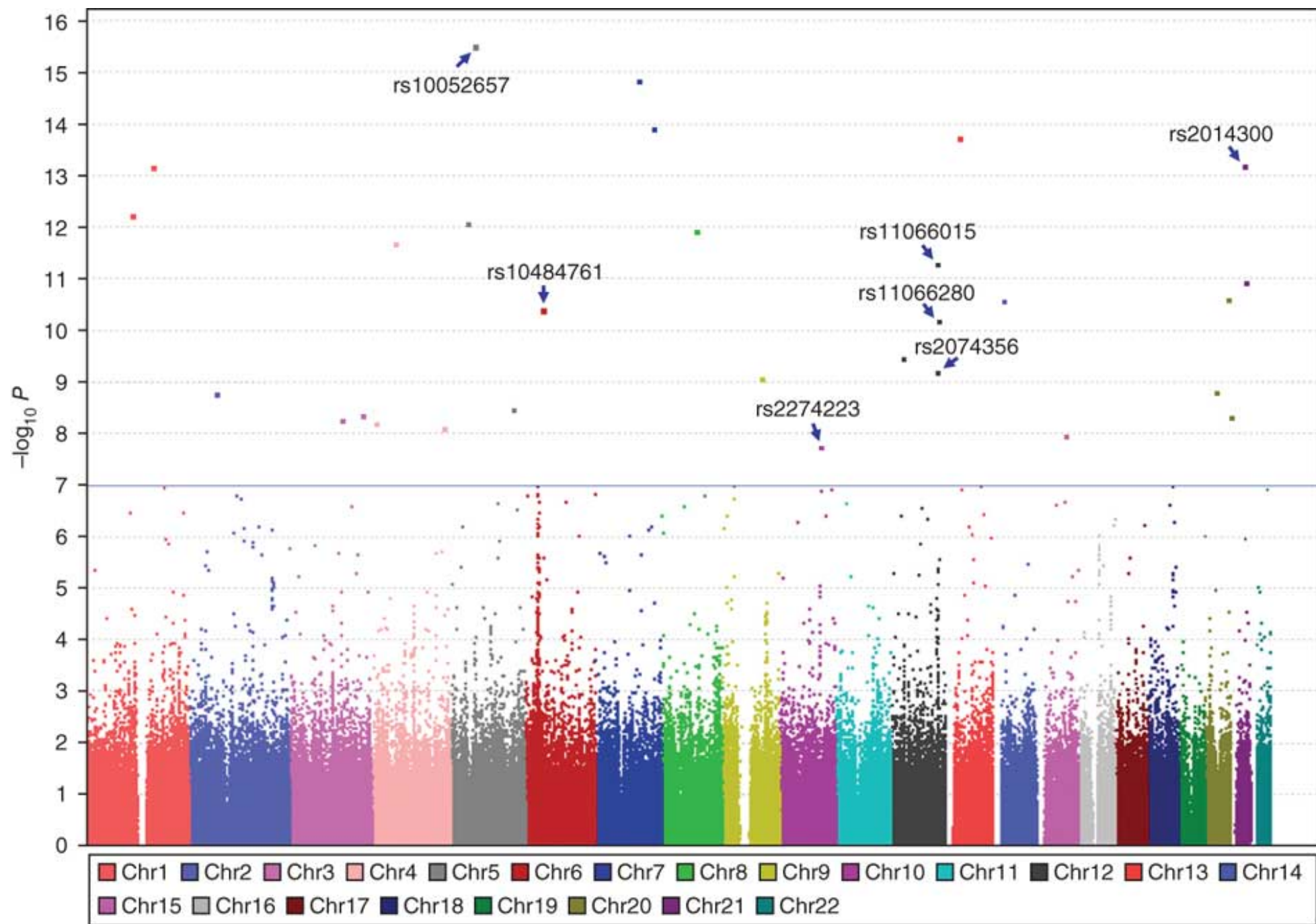
- If one has a very large number of SNPs, then new mutations (such as those that influence a trait) will be in **LD with very close SNPs** for hundreds to thousands of generations, generating a marker-trait association.
 - Association mapping looks over all sets of SNPs for trait-SNP associations. GWAS = genome-wide association studies.
 - This is also the basis for genomic selection
- Main point from extensive human association studies
 - Almost all QTLs have very small effects
 - Marker-trait associations do not fully recapture all of the additive variance in the trait (due to incomplete LD)
 - This has been called the “missing heritability problem” by human geneticists, but not really a problem at all (more shortly).

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



Population Stratification

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

$$\text{Cov}(M,y) \neq 0$$

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

Accounting for population structure

- Three classes of approaches proposed
 - 1) Attempts to correct for common pop structure signal (**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.

Regression Approaches

One 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

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

SNP marker
under consideration

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

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.

Mixed-model approaches

- Mixed models use marker data to
 - Account for population structure
 - Account for cryptic relatedness
- Three general approaches:
 - Treat a single SNP as fixed
 - TASSLE, EMMA
 - Treat a single SNP as random
 - General pedigree method
 - Fit all of the SNPs at once as random
 - GBLUP

Structure plus Kinship Methods

Association mapping in plants often 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 (fits SNPs one at a time)

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.

$\text{Cov}(u) = \text{var}(A) * A$, 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

Treat Single SNP as random: General Pedigree method

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

Here

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

Estimated from marker
data

Estimated from
the pedigree

The resulting likelihood function is

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

A significant σ_A^2 indicates a linked QTL.

GBLUP

- The Q-K method tests SNPs one at a time, treating them as fixed effects
- The general pedigree method (slides 24-26) 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., $\frac{1}{2}$ 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 (or K or marker matrix M)
 - Hence “identical” relatives can differ significantly in fraction of shared regions
 - Dense marker information can account for this

The general setting

- Suppose we have n measured individuals (the $n \times 1$ vector \mathbf{y} of trait values)
- The $n \times n$ relationship matrix \mathbf{A} gives the relatedness among the sampled individuals, where the elements of \mathbf{A} are obtained from the pedigree of measured individuals
- We may also have p ($\gg n$) SNPs per individual, where the $n \times p$ marker information matrix \mathbf{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 \mathbf{u} of random effects
- Standard form is that $\text{Cov}(\mathbf{u}) = \text{variance component} * \text{matrix of known constants}$
 - This is the case for pedigree data, where \mathbf{u} is typically the vector of breeding values, and the pedigree defines a **relationship matrix** \mathbf{A} , with $\text{Cov}(\mathbf{u}) = \text{Var}(A) * \mathbf{A}$, the additive variance times the relationship matrix
 - With marker data, the covariance of random effects are functions of the marker information matrix \mathbf{M} .
 - If \mathbf{u} is the vector of p marker effects, then $\text{Cov}(\mathbf{u}) = \text{Var}(m) * \mathbf{M}^T\mathbf{M}$, the marker variance times the covariance structure of the markers.

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

Pedigree-based BV estimation: (BLUP)

$u_{n \times 1}$ = vector of BVs, $\text{Cov}(u) = \text{Var}(A) A_{n \times n}$

Marker-based BV estimation: (GBLUP)

$u_{n \times 1}$ = vector of BVs, $\text{Cov}(u) = \text{Var}(m) M^T M$ ($n \times n$)

GWAS: $u_{p \times 1}$ = vector of marker effects,

$\text{Cov}(u) = \text{Var}(m) M M^T$ ($p \times p$)

Genomic selection: predicted vector of breeding values from marker effects, $GBV_{n \times 1} = M_{n \times p} u_{p \times 1}$.

Note that $\text{Cov}(GBV) = \text{Var}(m) M^T M$ ($n \times n$)

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

GWAS Model diagnostics

The “Genomic Control” parameter λ

Devlin and Roeder (1999). Basic idea is that association tests (marker presence/absence vs. trait presence/absence) is typically done with a standard 2×2 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 $S \sim \lambda\chi^2_1$). Hence, population structure should inflate all of the tests (on average) by a common amount λ .

Hence, if we have suitably corrected for population structure, the estimated inflation factor λ among tests should be ~ 1 .

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

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

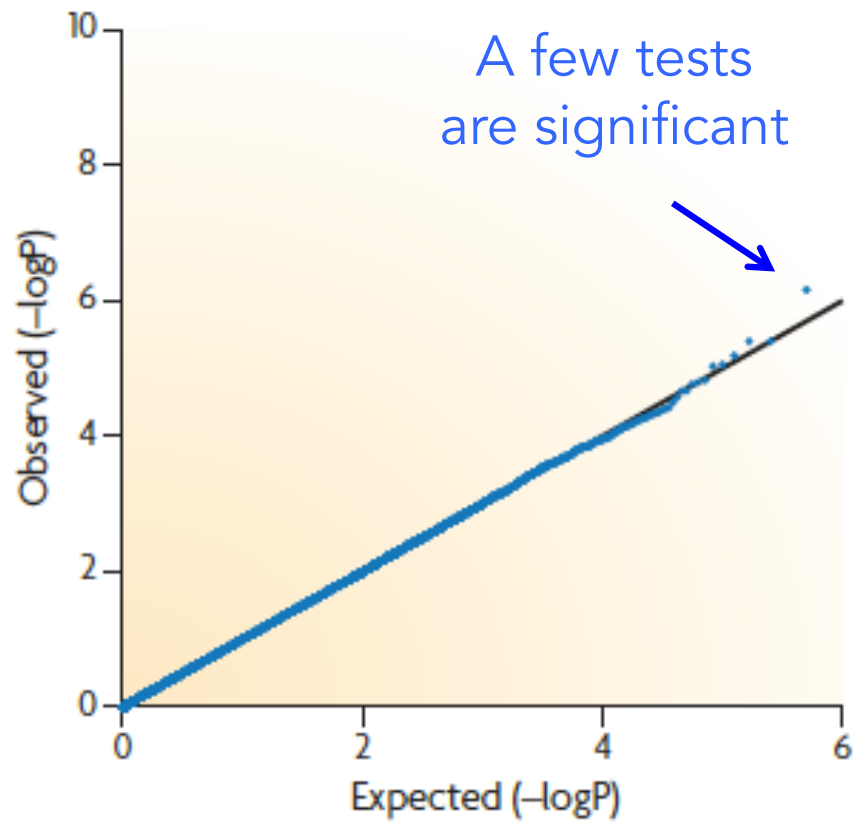
Genomic control λ as a diagnostic tool

- 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 than 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.

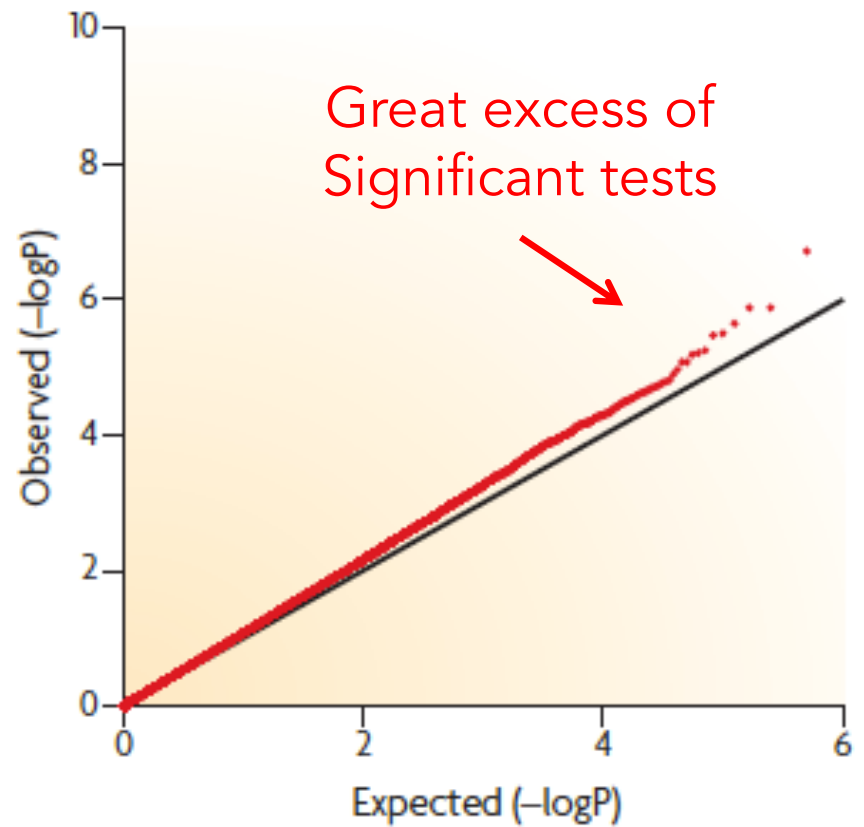
P – P plots

- 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

a No stratification

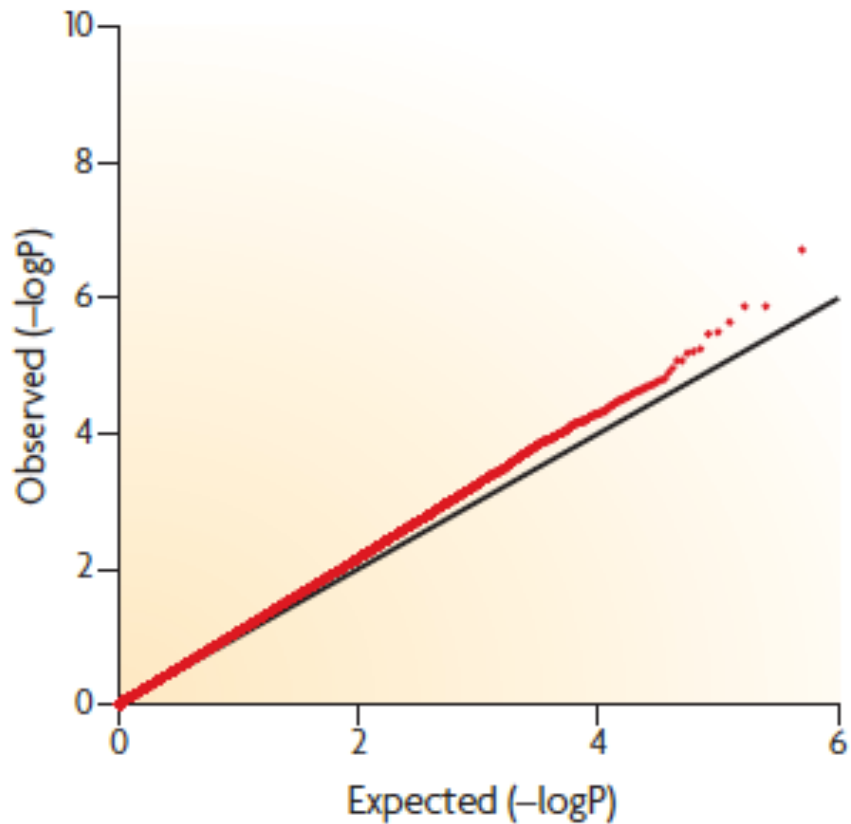


b Stratification without unusually differentiated markers

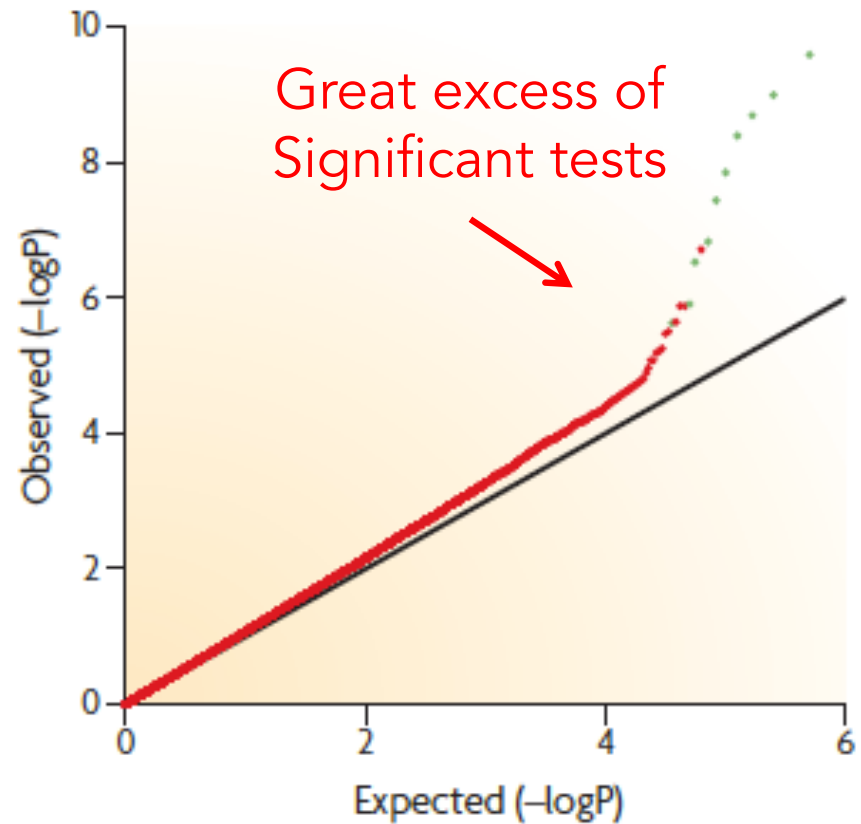


Price et al. 2010 Nat Rev Gene 11: 459

b Stratification without unusually differentiated markers



c Stratification with unusually differentiated markers



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

Association mapping (power)

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 $Q = 2(rp)(1-rp)a^2$
 $\sim 2rpa^2$ when Q rare. Hence, little power if Q rare

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

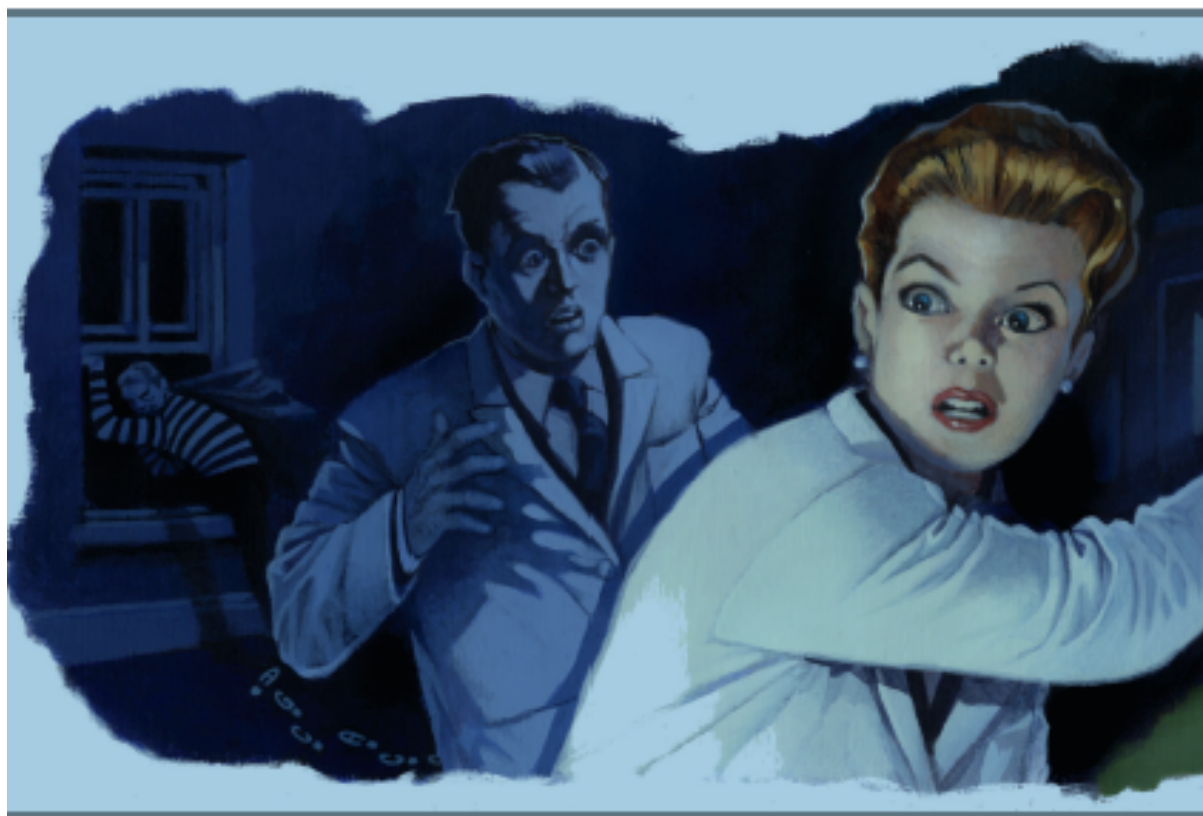
Ratio of marker/true effect variance is $\sim r$

Hence, if Q rare within the A class, even less power, as M only captures a fraction of the associated QTL.

Common variants

- Association mapping is only powerful for **common variants**
 - $\text{freq}(Q)$ moderate
 - $\text{freq}(r)$ of Q within M haplotypes modest to large
- Large effect alleles (a large) can leave small signals.
- The fraction of the actual variance accounted for by the markers is no greater than $\sim \text{ave}(r)$, the average frequency of Q within a haplotype class
- Hence, don't expect to capture all of $\text{Var}(A)$ with markers, esp. when QTL alleles are rare but markers are common (e.g. common SNPs, $p > 0.05$)
- Low power to detect $G \times G$, $G \times E$ interactions

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



The case of the missing heritability

Infamous figure from *Nature* on the angst of human geneticists over the finding that all of their discovered SNPs still accounted for only a fraction of relative-based heritability estimates of human disease.

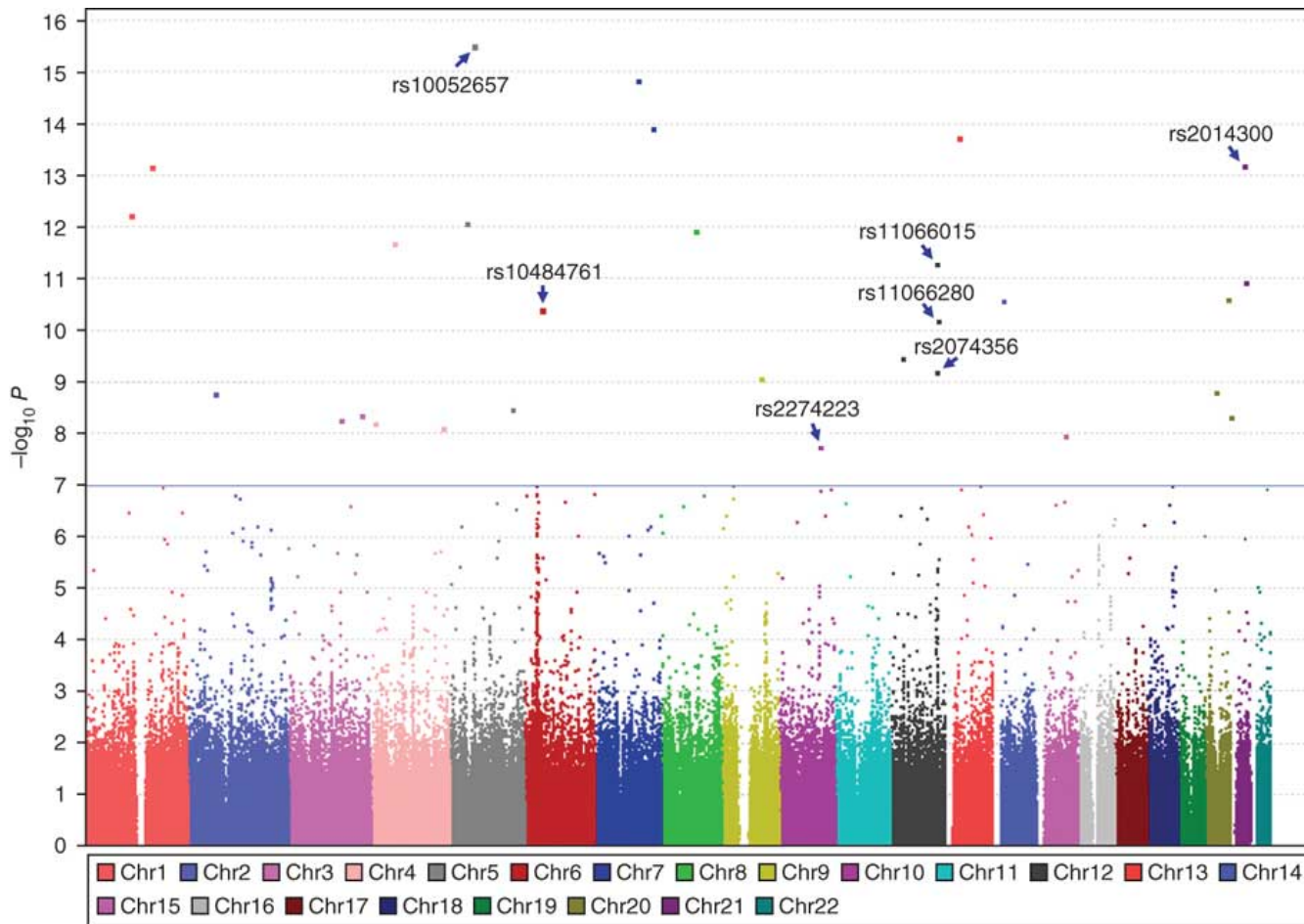
- “There is something simultaneously remarkable and encouraging about the fact that a centuries-old method requiring no more than a ruler, a pencil and (I suppose) a slide rule outperformed, by an order of magnitude, the fruits of the genomic revolution”
- --Ben Sheldon (2013)

The “missing heritability” 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 $\sim r$ (the frequency of the causative site within a marker haplotype class) of the underlying variance is accounted for.

No "missing heritability"

- **Low power** because sites of small effect are unlikely to be called as significant, esp. given the high stringency associated with control of false positives over tens of thousands of tests.



Only these markers
Included (as they are
declared significant)

Huge number of important,
but small effect, markers
not declared significant

Further, even with
all markers included,
only a fraction r of
variation explained