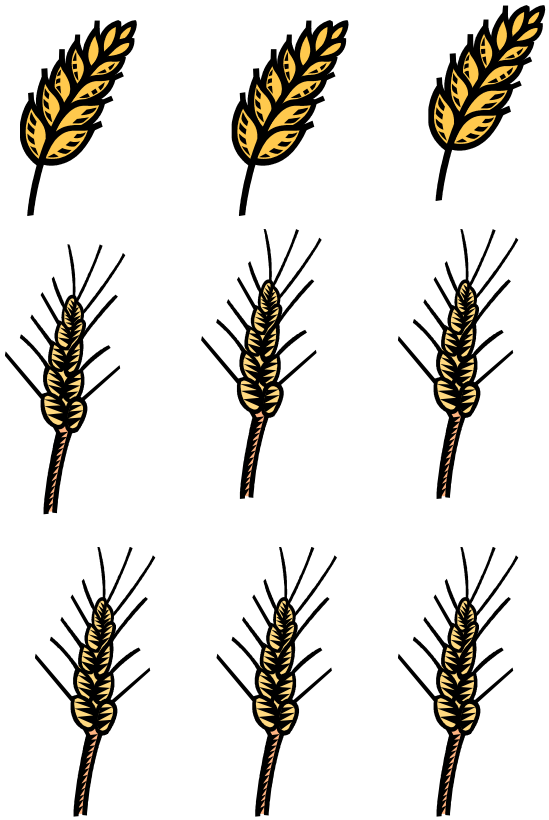
Overview

- Association mapping
  - Problems
  - Statistical solutions
- Comments on design and power

Trait mapping using association

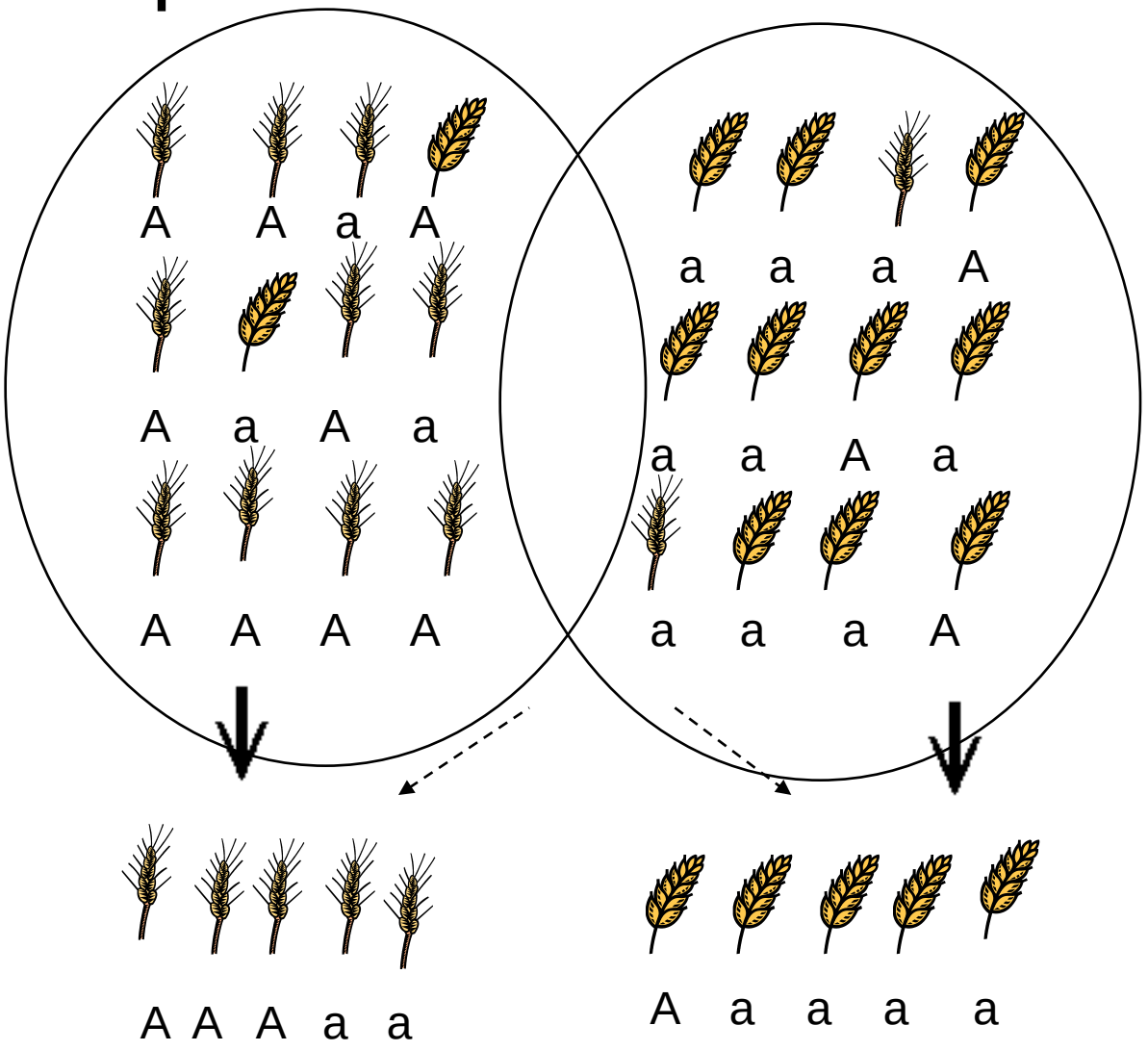
Allele A



Allele a

Allele a is found more frequently with 

Population structure and association



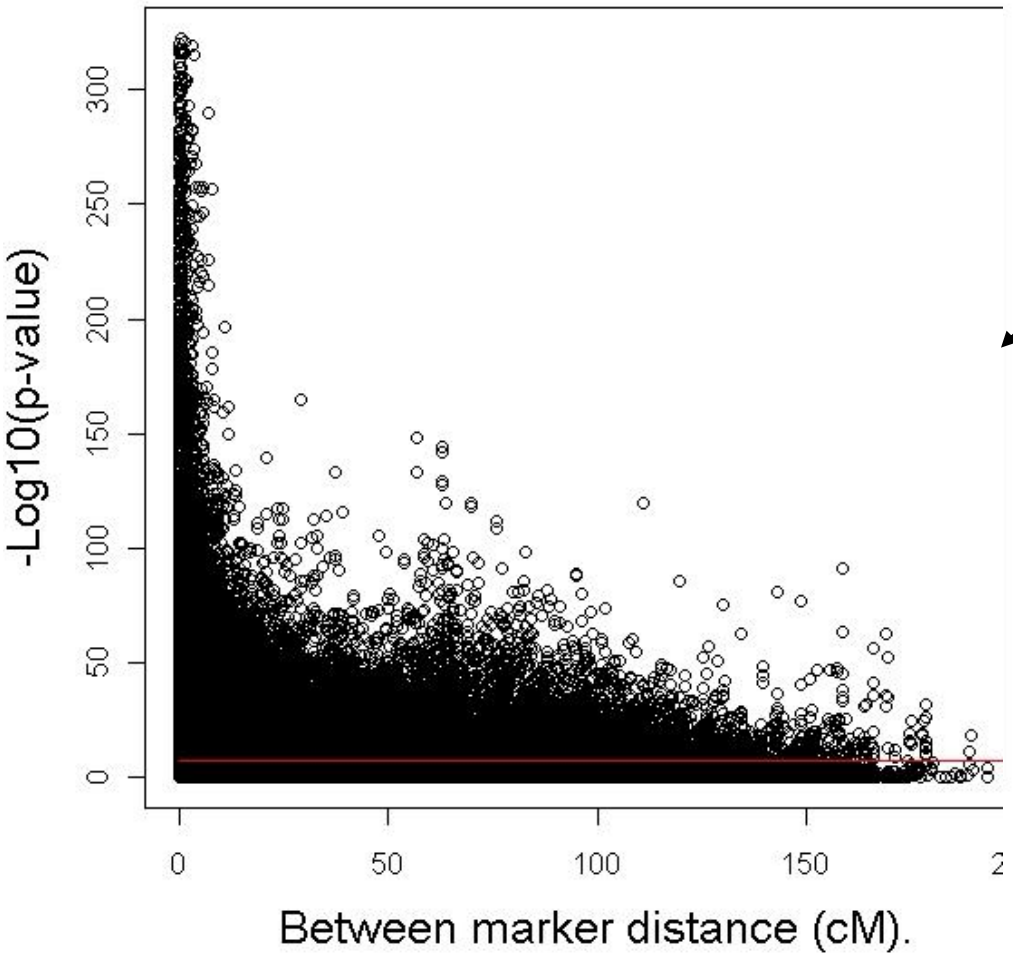
If there are unknown subgroups or families,

if allele freqs differ between subgroups,

if traits differ between subgroups,

then:
spurious association will be observed.

Between marker association falls with rising genetic distance.



Association without linkage.

Jon White, 15 March, 2011

LD mapping

Family based linkage mapping and LD mapping compared:

LD mapping exploits historic recombination in wild populations and is best at fine mapping.

Linkage analysis exploits contemporary recombination in experimental populations and is best at QTL detection.

Strengths and weaknesses of LD mapping

Kinship & pop structure	largely solved
Low power	need large pop sizes
Better precision	LD decays more rapidly
Use of existing data	historical collections
Need high marker density	will be solved

Pedigree structure generates false +ve's

1) Close kinship

Amounts to double counting: you have less data than you think.

2) Distant branches diverge: selection /drift /founder effects

Genotypes and phenotypes can differ between branches, causing associations across the genome at multiple loci.

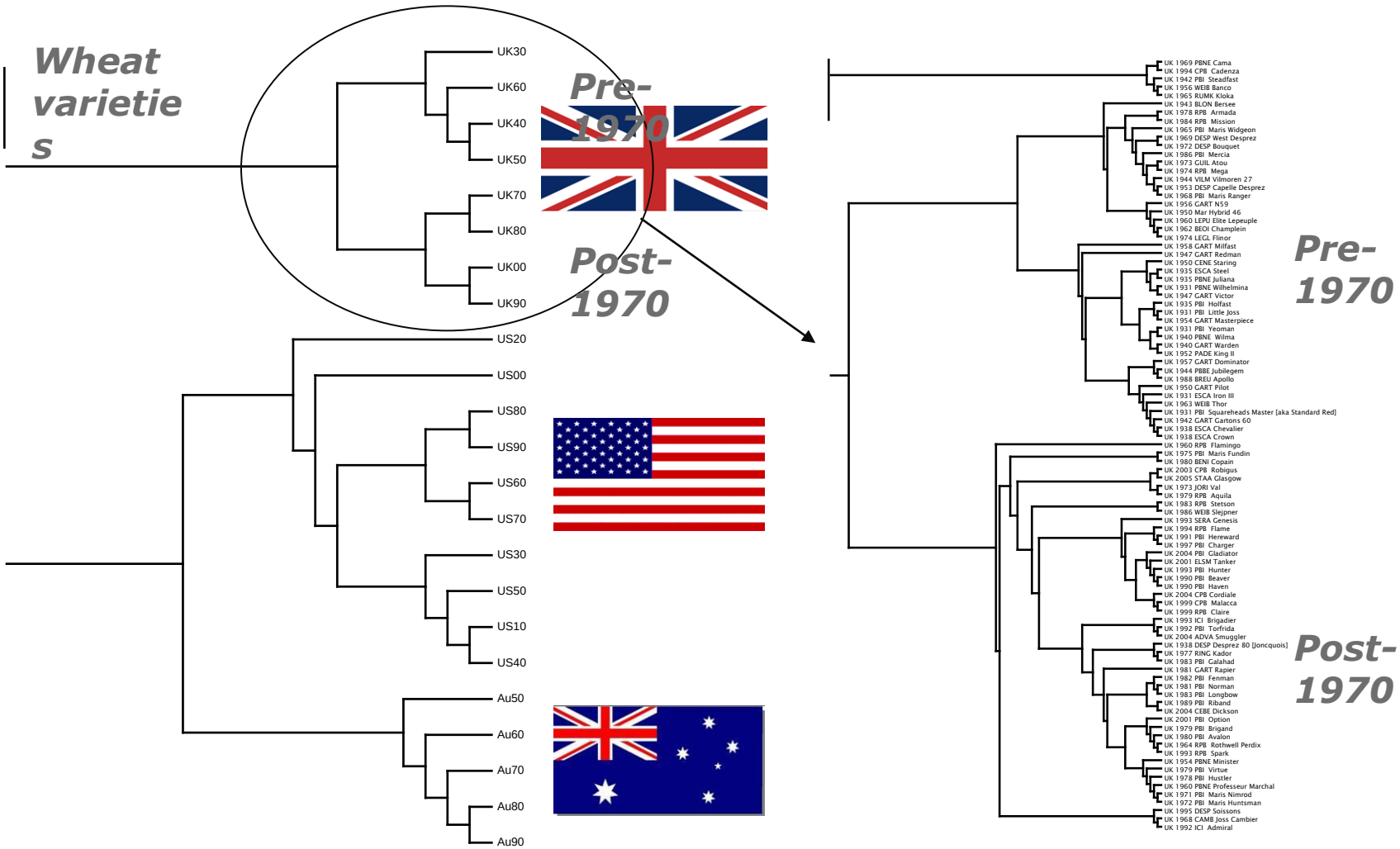
Any natural population will comprise a mixture of these effects.
Relative importance will vary with dataset.

Need to account for both.

In crops, problems associated with kinship effects are massive.

Not a problem in experimental mapping populations eg an F2

Kinship or Structure?



Jon White, 15 March, 2011

Experimental solutions

The transmission disequilibrium test (also QTDT, PTD etc.)		humans
Nested Association mapping	-	maize (Buckler)
MAGIC	-	mouse, Arabidopsis, wheat
Selection experiments		

Analytical solutions

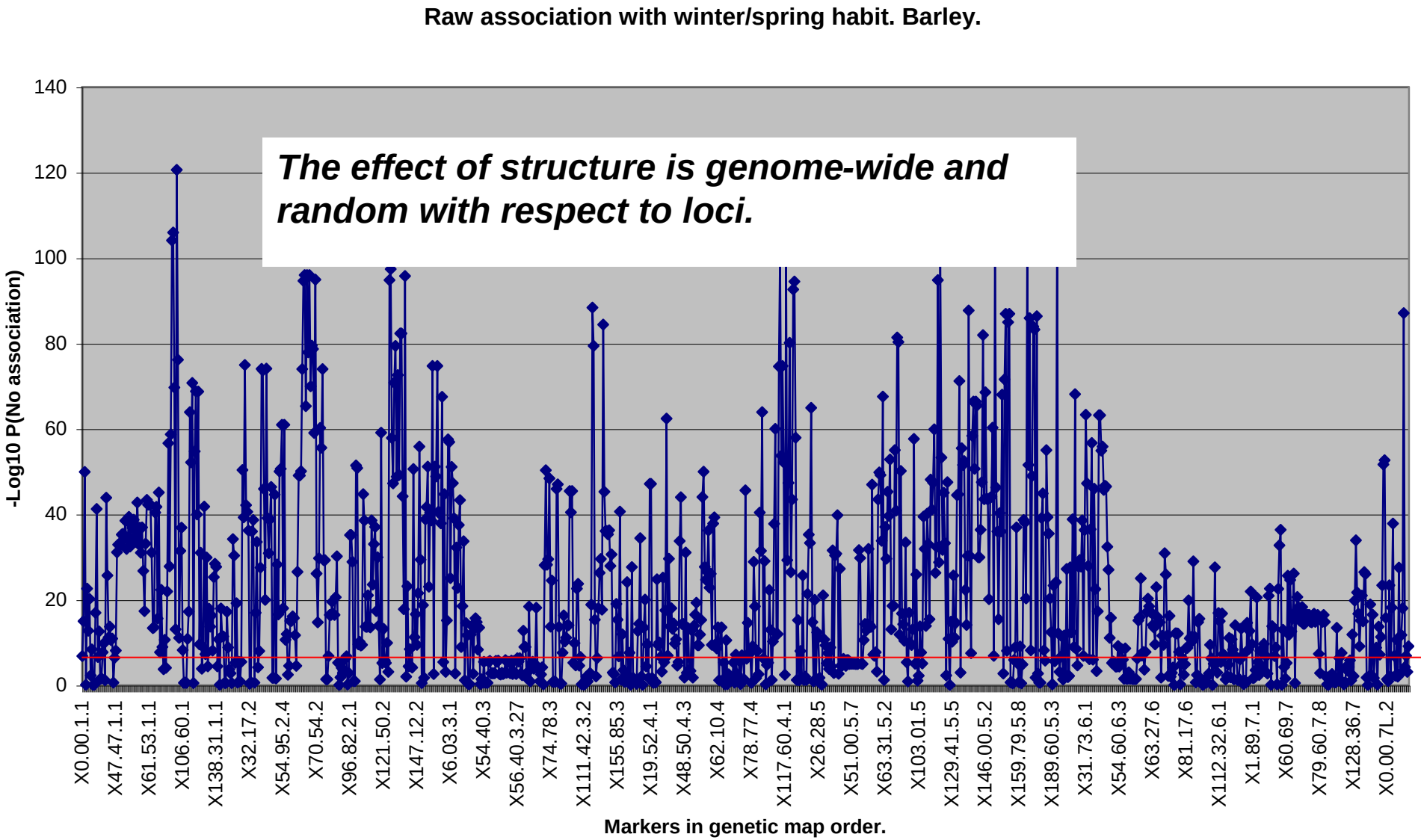
Genomic control: returning the mean of the distribution of the test statistic to its expectation under the null.

Structured Association: simple linear regression but include covariates to account for subpopulation membership. Use *STRUCTURE* to get the covariates

PCA: Similar to SA but use PCA to adjust both and phenotype for subpopulation membership in terms of top (20) eigenvectors of the correlation matrix; measure association in terms of correlation between the residuals from these models.

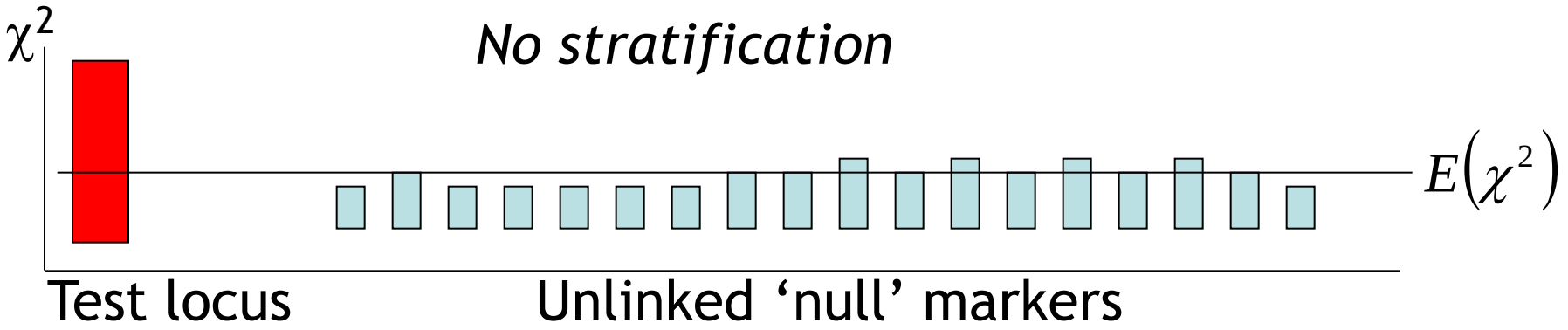
Mixed Model: currently the method of choice

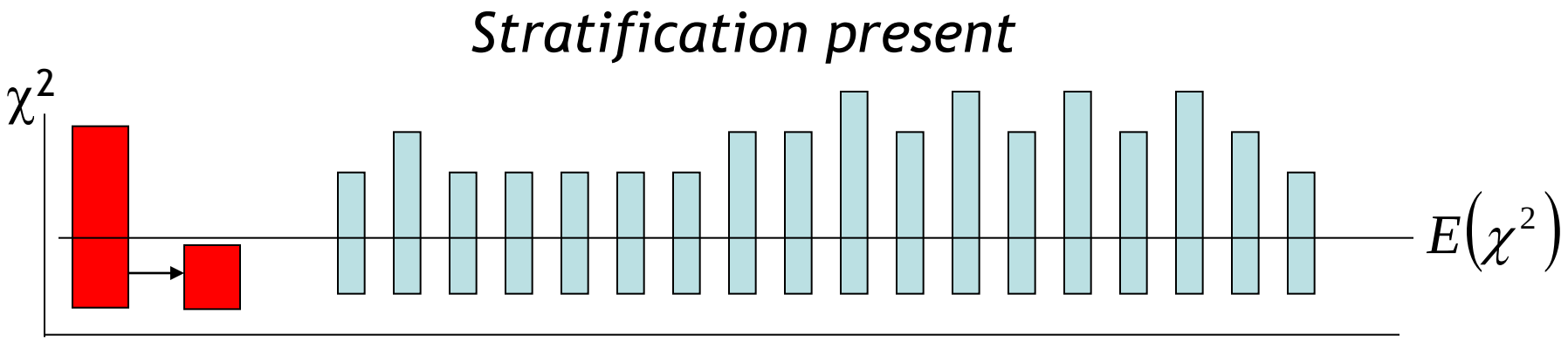
Others



Jon White, 15 March, 2011

Genomic control





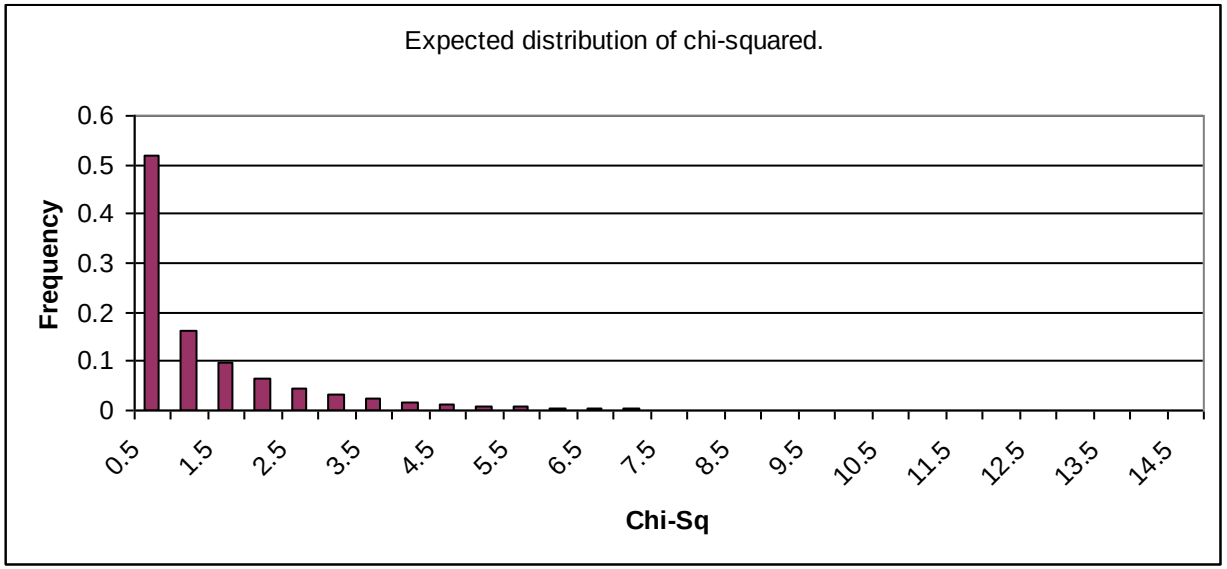
Stratification → adjust test statistic

NIAB

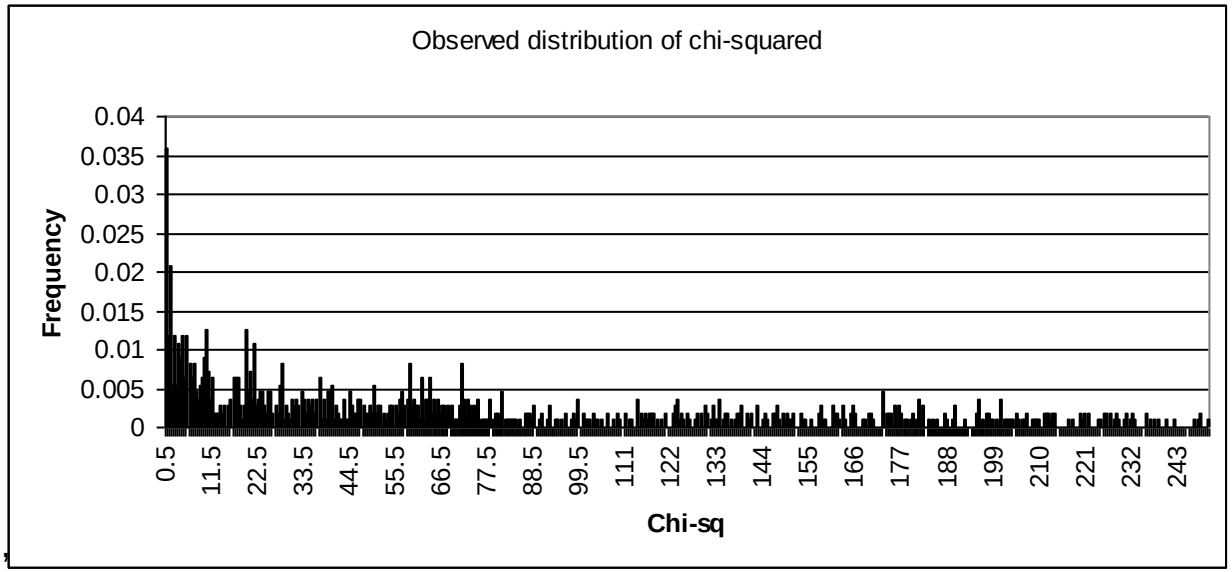
Genomic Control: Rationale

- There is association at nearly all markers.
- We know the expected distribution of the test statistic under the null hypothesis.
- We really only expect association at a few markers.
- So we would not expect the observed distribution to be much different from the expected.

Genomic control: Example using Chi-Squared, 1 d.f.



Mean: 1.0
Median: 0.456



Mean: 94.07
Median: 57.28

Jon White, 2011

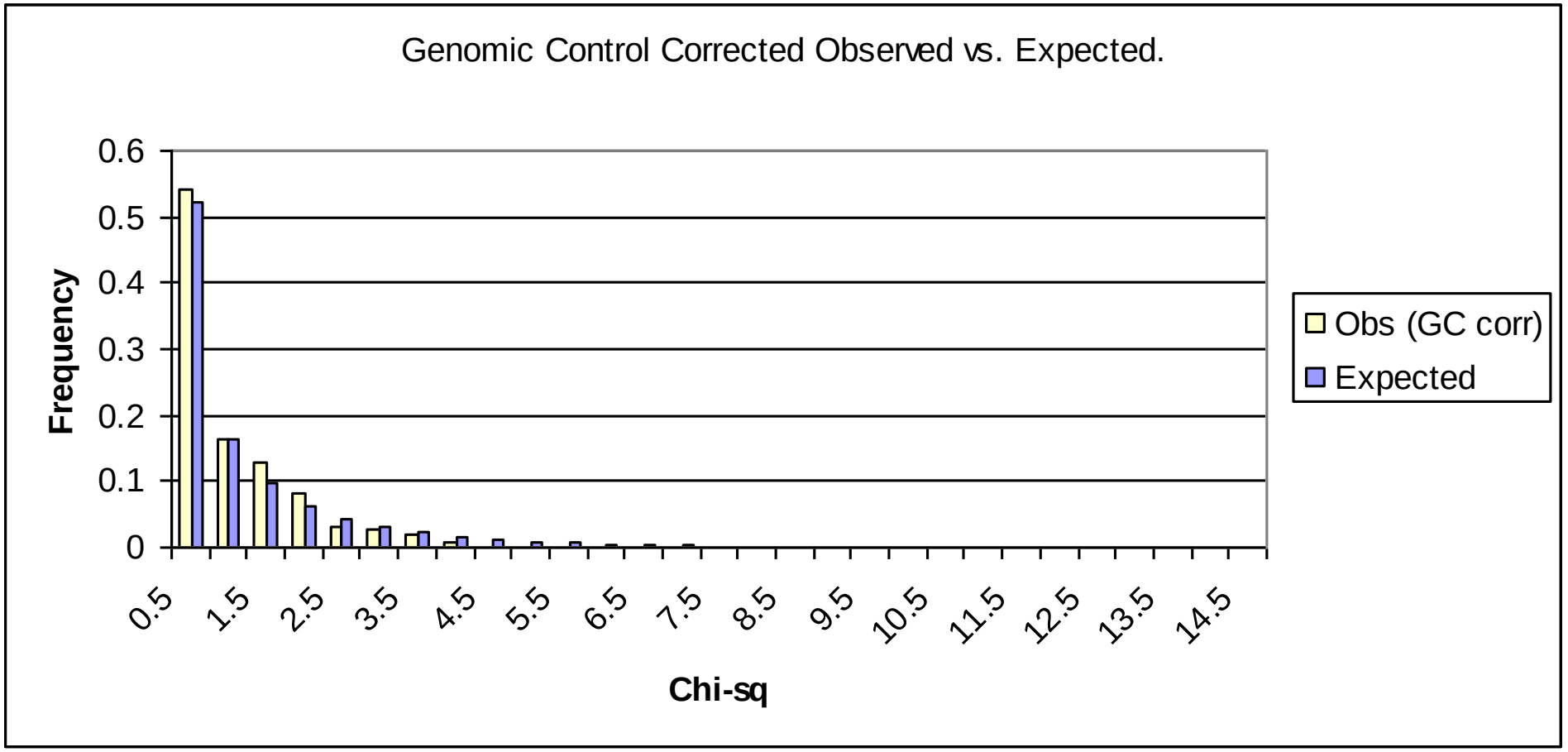
Genomic Control

$$CorrectedChiSq = \frac{ObservedChiSq}{ObservedMedianChiSq} \times ExpectedMedianChiSq$$

In our example:

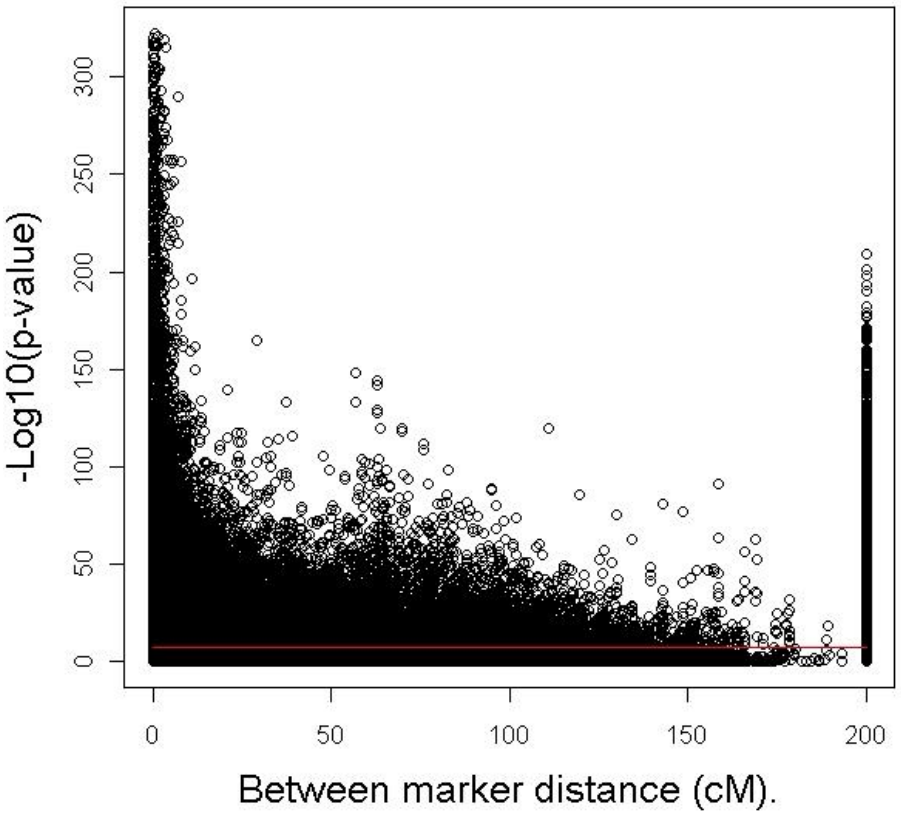
$$CorrectedChiSq = \frac{ObservedChiSq}{57.28} \times 0.456$$

Some authors correct using observed and expected mean.

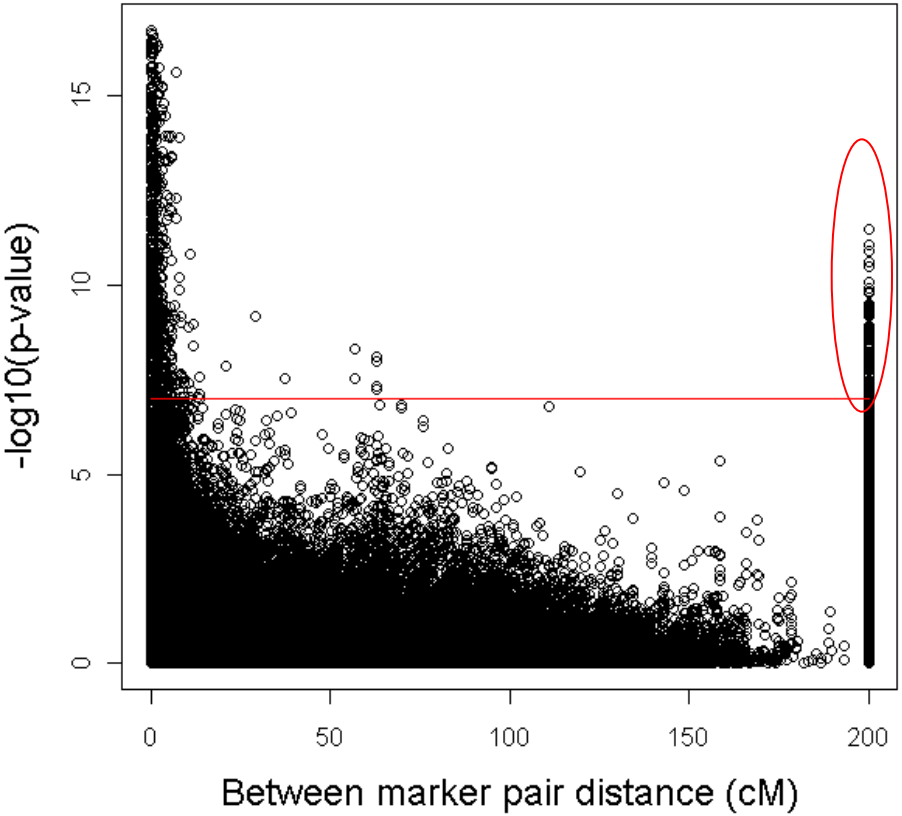


Jon White, 15 March, 2011

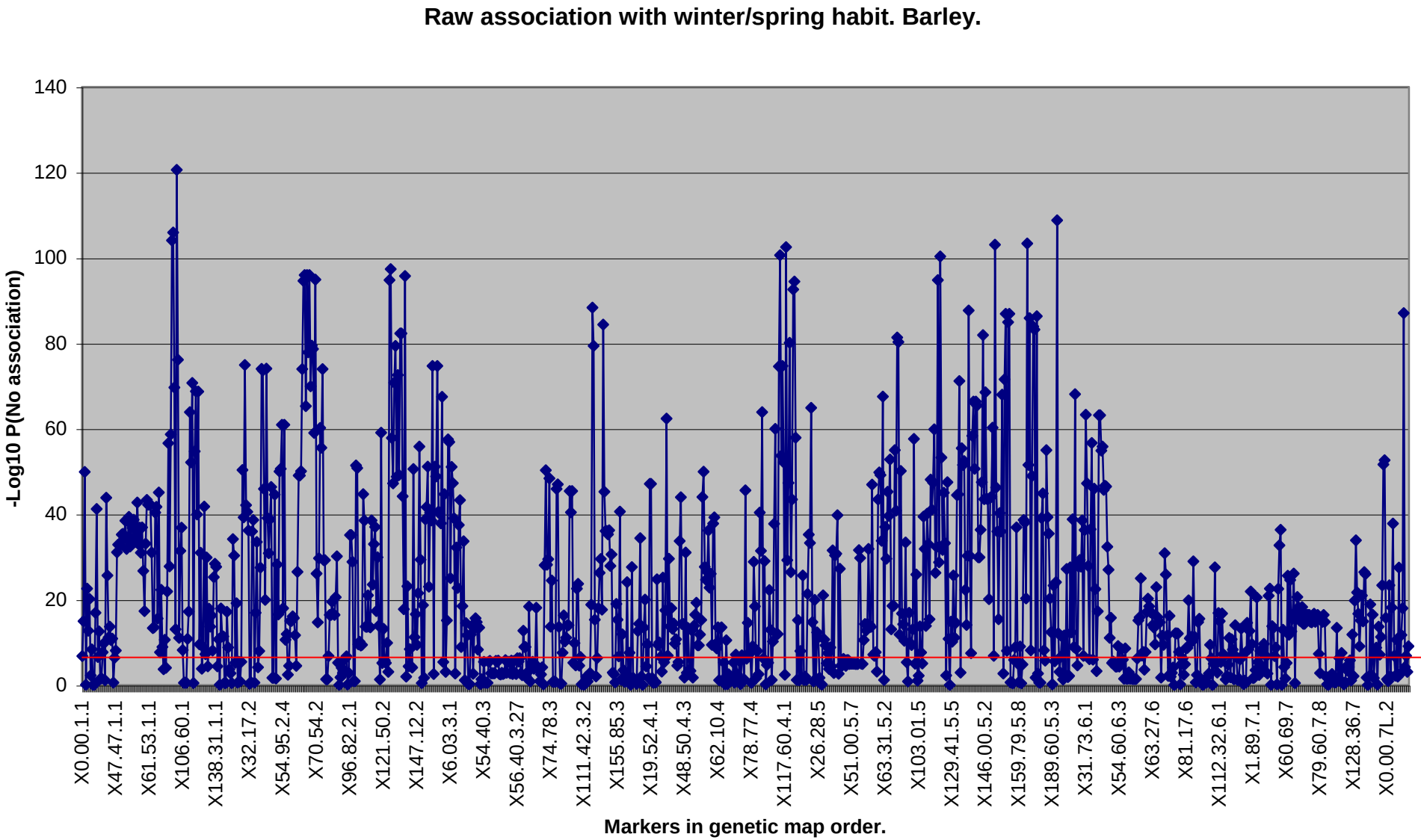
Before GC



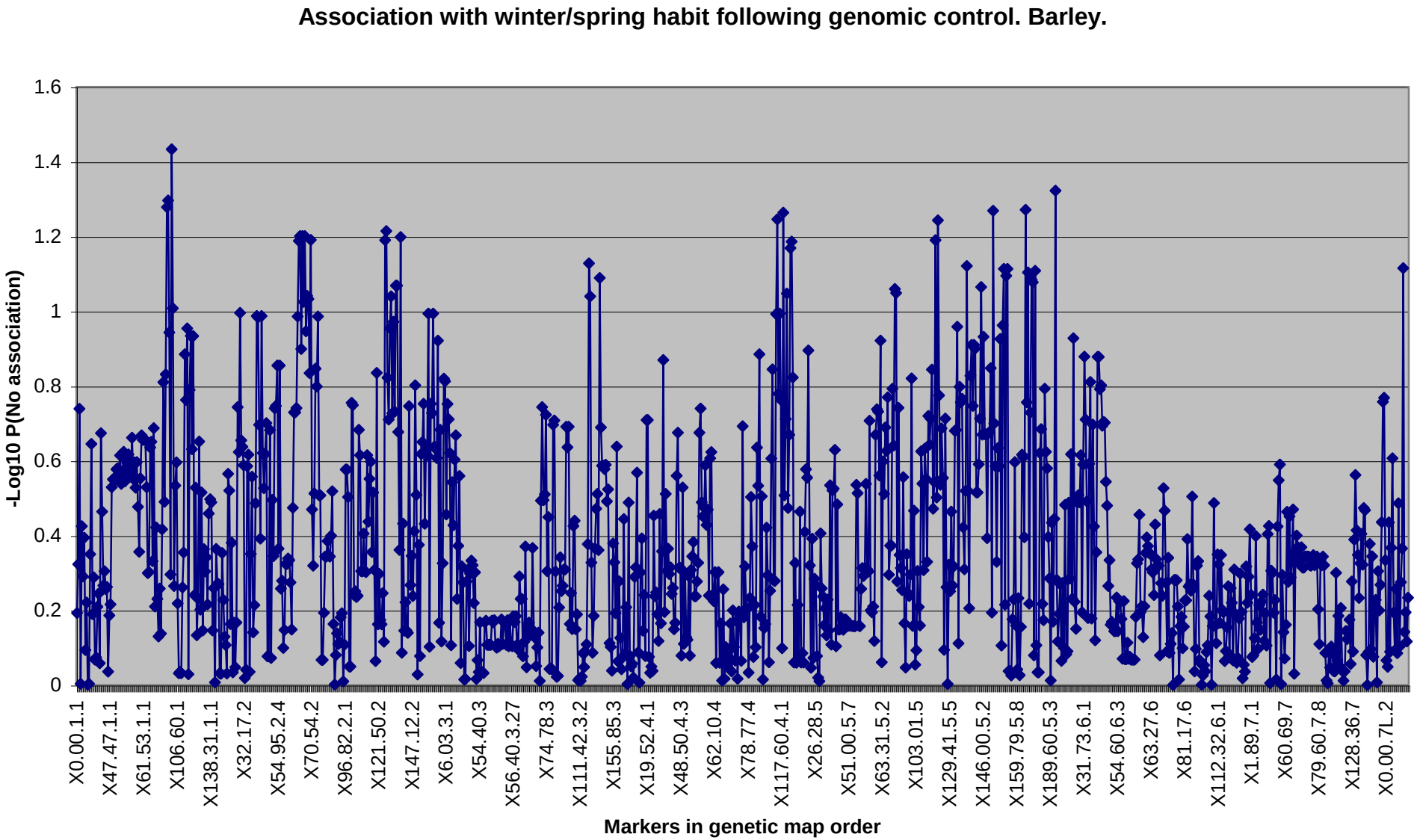
After GC



Jon White, 15 March, 2011



Jon White, 15 March, 2011



Jon White, 15 March, 2011

Genomic Control

- Corrects the symptoms of structure.
- Does not change the ranking of significance of association.
- Loss of statistical power.

Key Benefit.

- *Returns the distribution of the test statistic close to its expectation: Allows us to work with conventional significance thresholds*.* *(*Well almost!)*

Structured Association

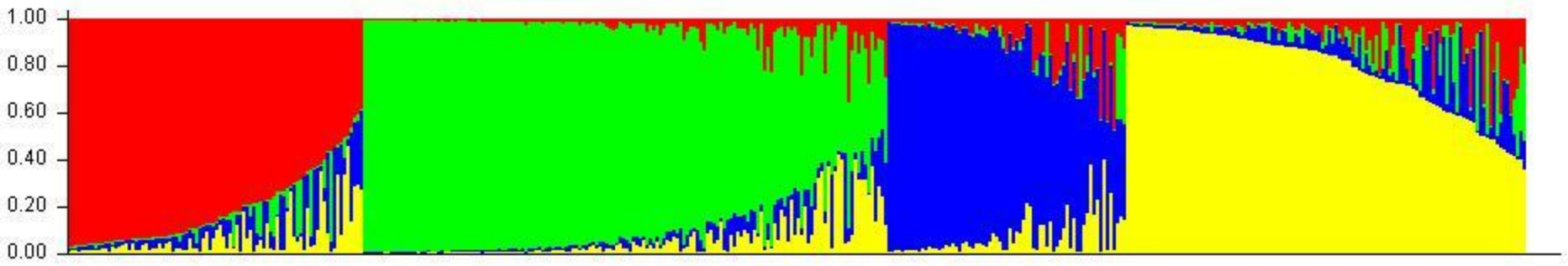
Estimate the ancestry of each individual in the sample. Most common is to use the programme Structure.

Regress the phenotype on the ancestry coefficients (to adjust for effects of population structure) and then on the test marker.

Does not correct adequately for recent coancestry – pedigree relationships.

Structure View

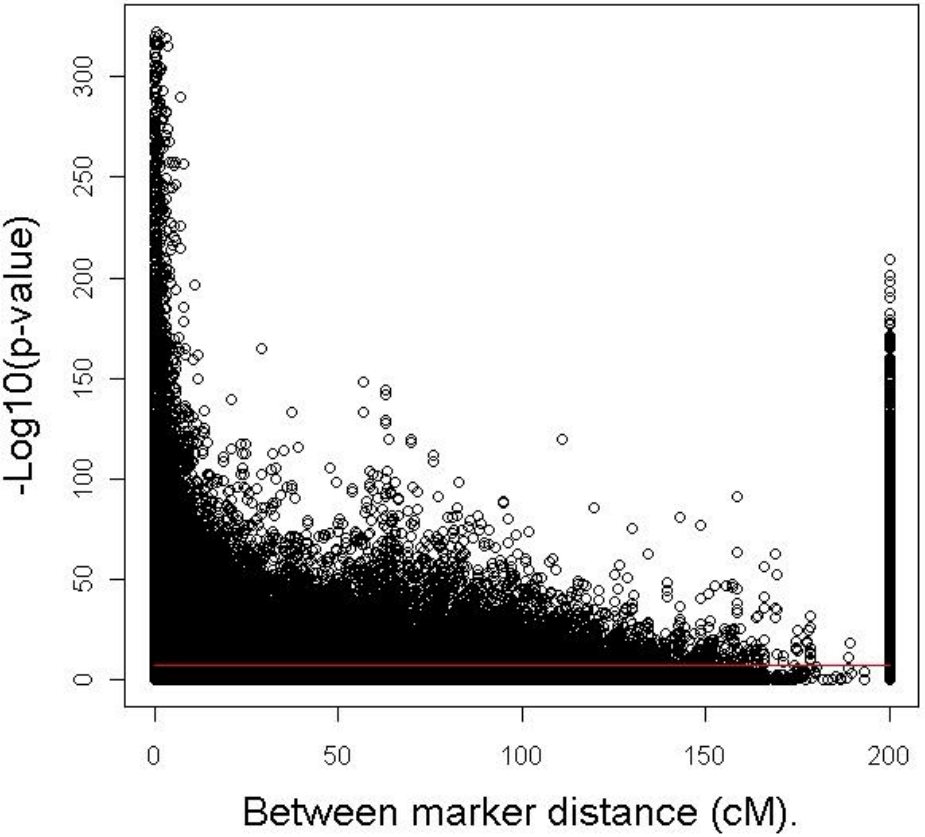
Software: Structure v2.2.

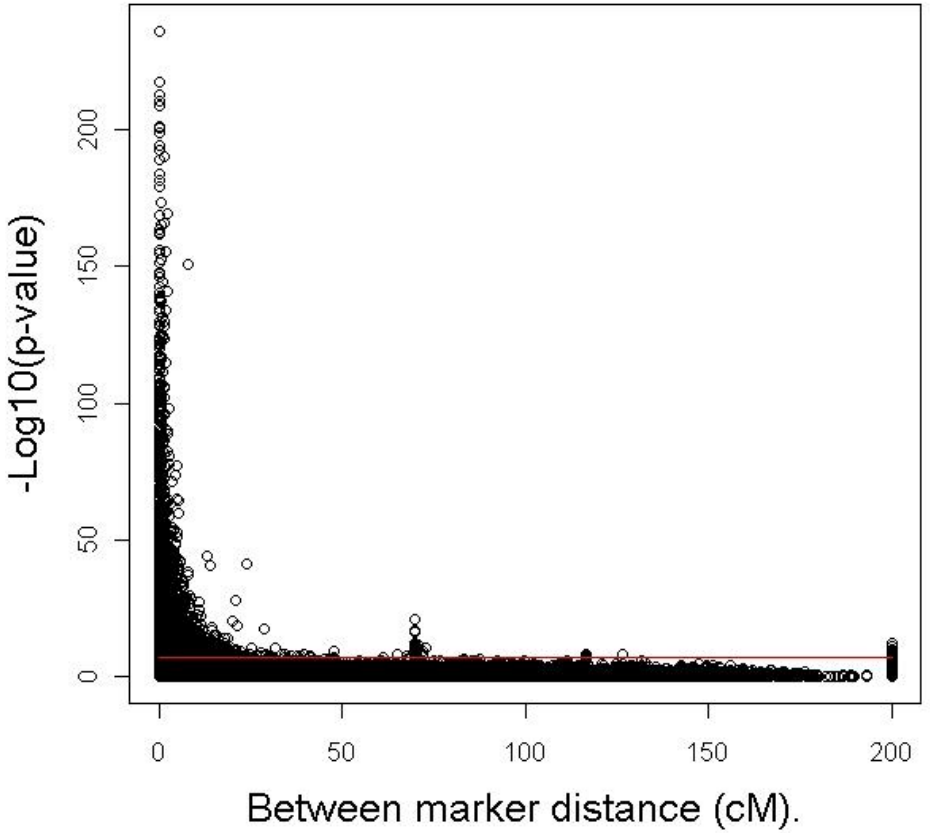


Q Matrix of Fractional Sub-population membership

Variety	K1	K2	K3	K4
A	0.1	0.0	0.0	0.9
B	0.5	0.0	0.0	0.5
C	0.9	0.1	0.0	0.0

Effect of structure specific correction for population structure





Bonferroni corrected. (P=0.05)

Principal component analysis

PCA of genotype data gives an indicator of ancestry for each variety (the eigenvector) for a population characterised by its the eigenvalue.

The deviation for each variety from a multiple regression of phenotype on eigenvectors gives a a new phenotype adjusted for population structure.

Deviations from regression of candidate markers on eigenvectors gives adjusted genotypes in the same way.

Correlation between adjusted phenotype and adjusted genotype is a a measure of association adjusted for the effect of population structure.

Advice is to include ~ 20 largest principle components for ancestry

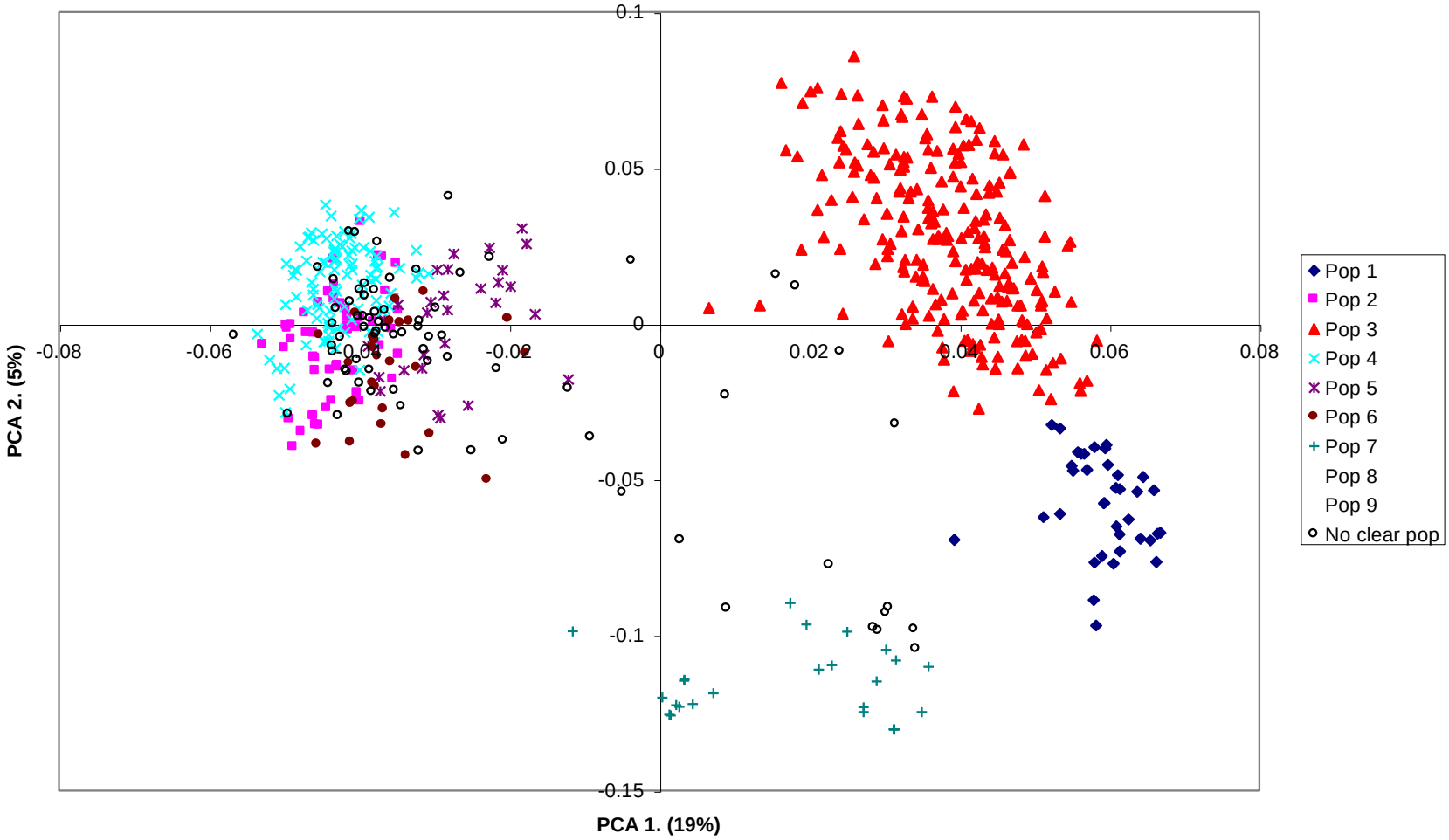
Currently only works for bi-allelic markers.

Will not adjust for recent coancestry.

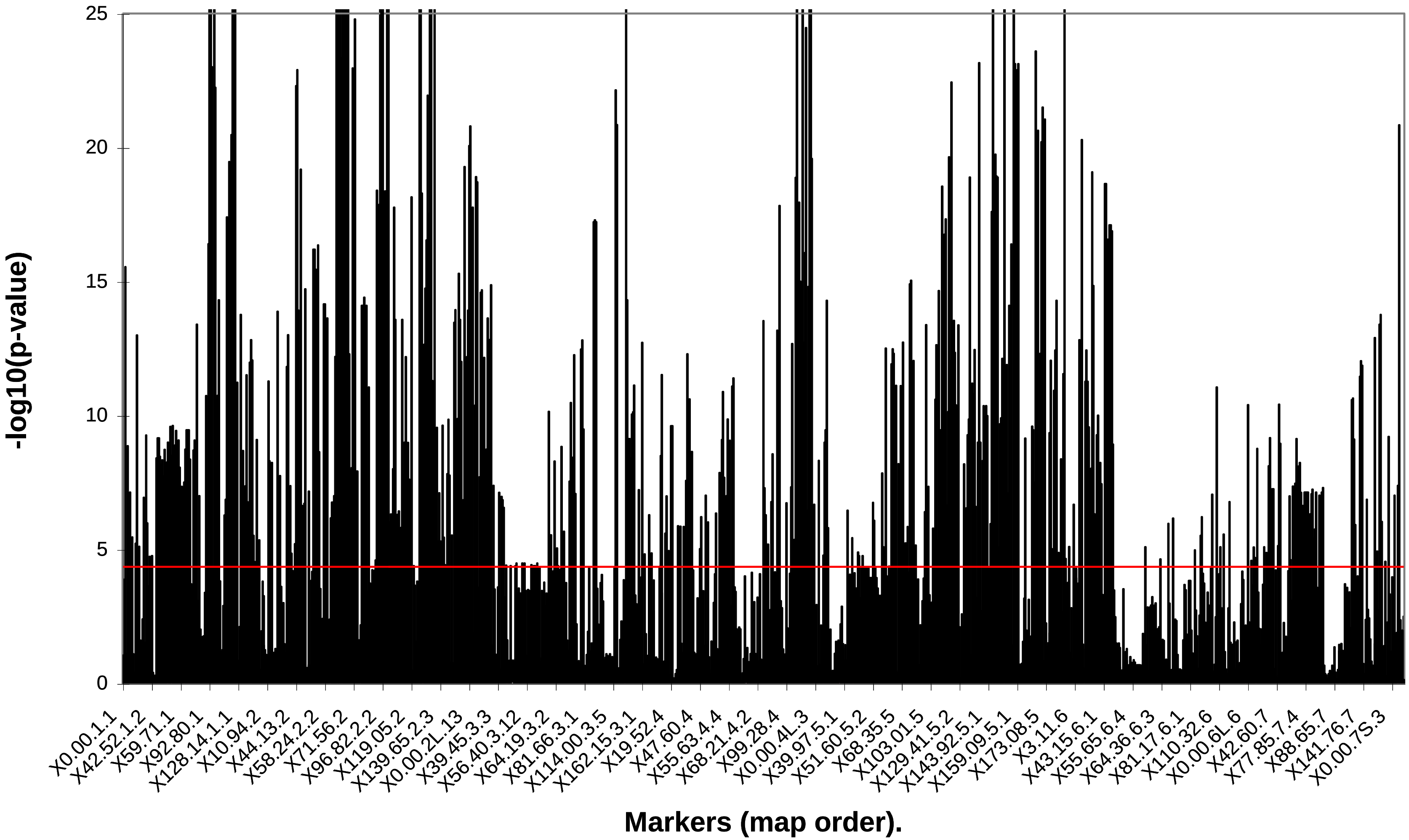
PCA based correction (a.k.a. Eigenstrat).

- Define the population structure in terms of co-variation between individuals.
- Simplify the information as principle components: eigenvectors.
- Use an informative subset of these vectors to predict genotype and phenotype.
- Calculate residuals
- Measure the correlation between the residual phenotype the residual genotype for each marker.

Principle components contain a lot of structural information.



Principle components used to correct for population structure.



Jon White, 15 March, 2011

Mixed models

Pedigree relationships mean that the error variances for each individual are no longer independent.

In addition to error variances, we must include in the model error covariances between related individuals.

If the relationships are known, these are fed into the model as expected genetic covariances among individuals. This is the basis of the mixed model. Software exists to do this automatically – GenStat, SAS, VCE and others.

If relationships are unknown, they can be estimated using markers, but these are not so easily fed into standard software which exploit properties of known pedigrees to greatly speed up computation. Use TASSEL, GenStat, EMMA, SAS

Mixed modelling adjusts for kinship yet still permits the inclusion of covariates to adjust for differences in phenotype between subgroups.

Mixed effects modelling

- Commonly implemented using software called TASSEL – we have found this very difficult to use.
- EMMA (Efficient mixed model analysis) is also available free and runs in R.
- A more rapid (and less temperamental) implementation of the EMMA method is soon to be available from Will Astle, Imperial College.

Strengths and weaknesses of LD mapping

Kinship & pop structure	largely solved
Low power	need large pop sizes
Better precision	LD decays more rapidly
Use of existing data	historical collections
Need high marker density	will be solved
Easy to publish “hits”	educate on good design

Want a cheap association mapping publication?

Do:
Use a small collection of cultivars.
Use a small number of “genome wide” markers.
Run STRUCTURE but with the default parameters.

Don’t:
Carry out power calculations.
Check for off-chromosome LD.
Check that “replicated QTLs” are no more than expected by chance.
Check the type 1 error rate.

Underpowered studies in crops, an exemplar:

Recently published:

7 chromosomes, 46 SSRs, 30 accessions.

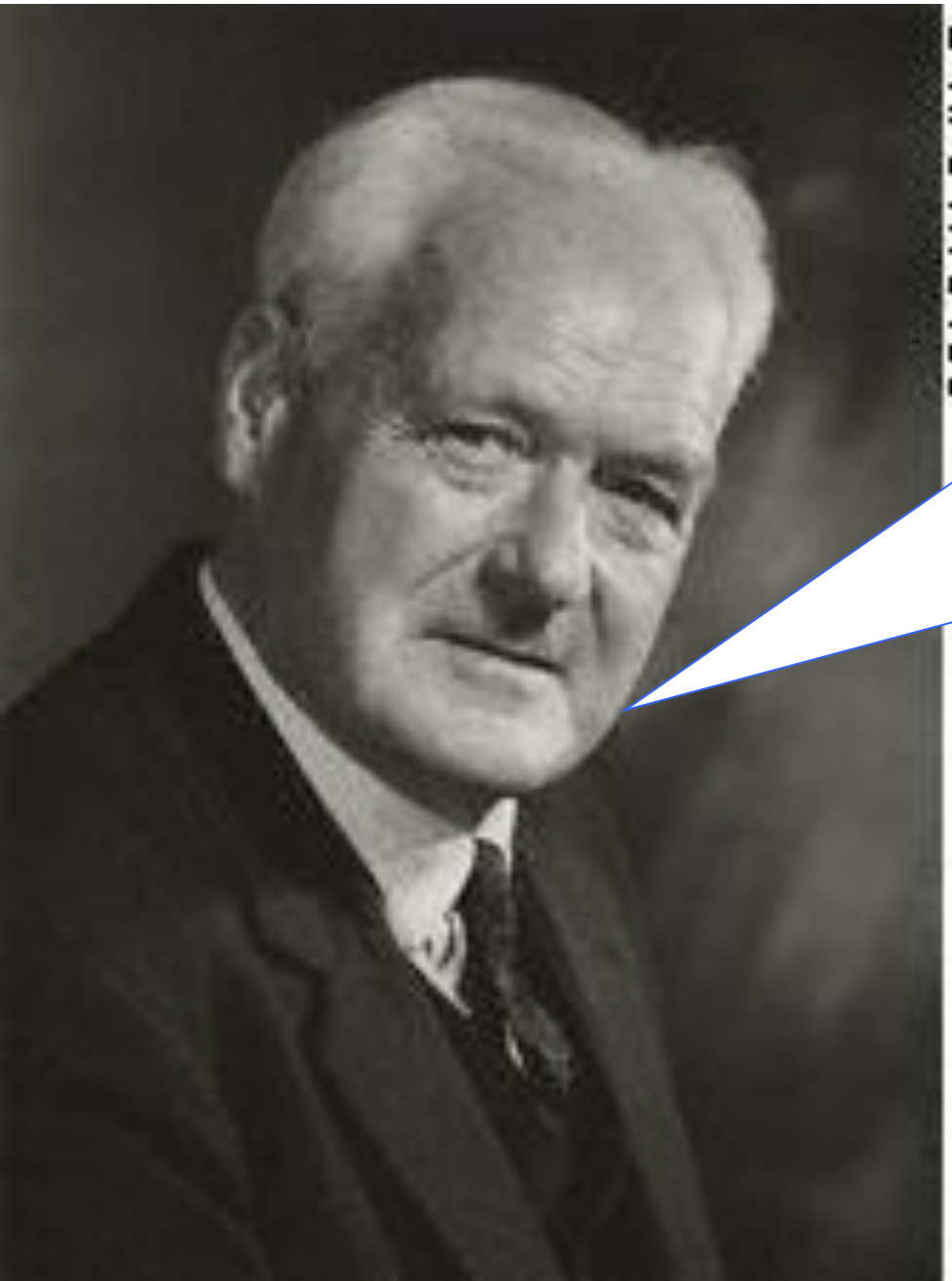
Multiple traits scored on 5 plants per accession.

No LD plots,
No power calculations,
Many positive results:

via screening only a small pool of germplasm [illegible] possible to identify a [illegible] [illegible] requires a large pool of germplasm to be screened in other species

“The present data also indicate that a few SSR markers were [illegible] efficiency of association mapping [illegible] resolution [illegible] is much higher in [illegible] than in other plant species.”

“The findings also indicate that the [illegible] association mapping [illegible]”



"The glitter of the *t* table diverts attention from the inadequacies of the fare."

Sir Austin Bradford Hill.

(First demonstrated the connection between smoking and lung cancer.)

Good study design is an ethical issue.

Extract from UK MREC application form:

13. Size of the study (including controls)

i) How many patients will be recruited?
ii) How many controls will be recruited?
iii) What is the primary end point?
iv) How was the size of the study determined?
v) What is the statistical power of the study?

A well designed association genetics study should consider:

marker density

LD decay

allele frequency distribution

number of samples

relatedness between samples

Are resources adequate for the objectives of the study?

What magnitude of effect are you likely to detect?

With what precision are you likely to locate QTL?

Are the results too good to be true?

Small studies can find major genes

10 cases, 10 controls,

(40 chromosomes)

recessive trait:

White spotting

Hair ridge.

Mapped to < 1 cM in ~ 20 dogs (40 chromsomes)

Nat Genet 2007 **39** p1321

2007: GWA studies come of age.

The Wellcome Trust Case Control Consortium, Nature 2007

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

24 genetic risk factors
Large collaborative effort: > 50 research groups
500 000 markers

But:

These explain only a small proportion of risk:
power is still low for the effect sizes in humans.

Hundreds of variants clustered in genomic loci and biological pathways affect human height.

Nature 2010 doi:10.1038/nature09410

h^2 ~80%

183,727 individuals

>180 loci, 100s of genetic variants

“Our data explain approximately 10% of the phenotypic variation in height”