The University of Newcastle

Kerrie Mengersen

Introduction to

Bayesian Methods for

QTL Analysis - 2

Armidale 2004

Bayesian Methods for QTL Analysis

This discussion is based on:

- Hoschele, I. Mapping Quantitative Trait Loci in Outbred Pedigrees
- Jansen, R. Quantitative Trait Loci in Inbred Lines. In Handbook of Statistical Genetics, Editors D.J. Balding, M. Bishop, C. Cannings. Wiley
- some slides by Charles Berry, ucsd.edu

Example 1: Inbred lines

- QTL: genes underlying quantitative or complex traits
- Consider first only inbred lines of diploid organisms (can generalise for polyploid organisms and biparental crosses between outbreeding lines).

Homozygous parents: P_1 (a_1a_1) P_2 (a_2a_2)
(identical alleles at any given gene)

Heterozygous offspring: F_1 (a_1a_2)

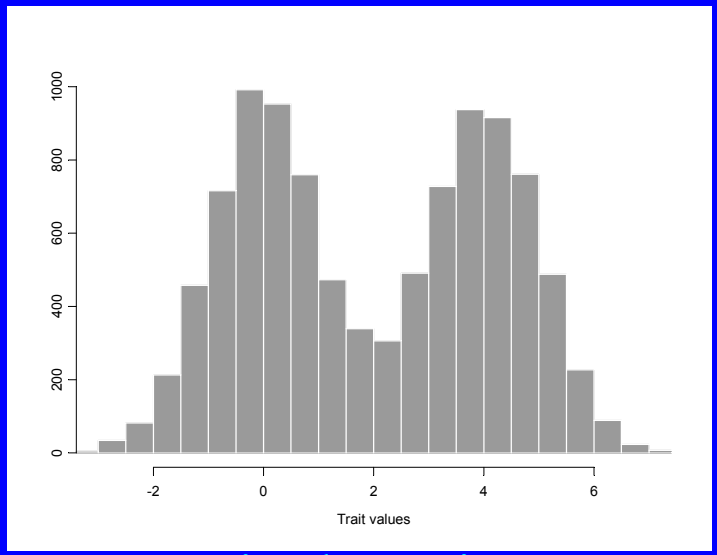
Backcross (BC) design: F_1 x P_1 or P_2 → eg F_1 x male P_1 (a_1a_2, a_1a_1)

Doubled haploids (DH) design: M or F gametes of F_1 artificially doubled (a_1a_1, a_2a_2)

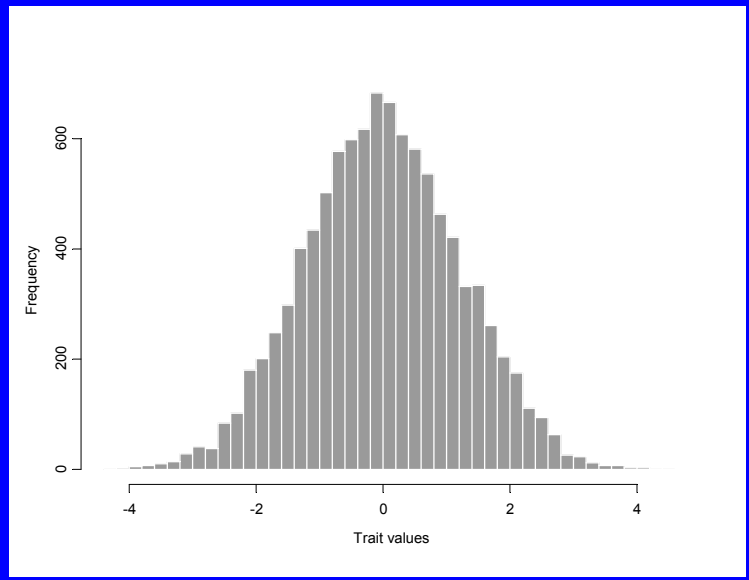
Filial F2 design: F_1 selfed, or 2 F_1 crossed (a_1a_1, a_1a_2, a_2a_2 ratio 1:2:1)

Recombinant inbred line (RIL): F2 progeny in single-seed descent inbreeding program (a_1a_2 x a_1a_2 → a_1a_1 or a_2a_2 per locus)

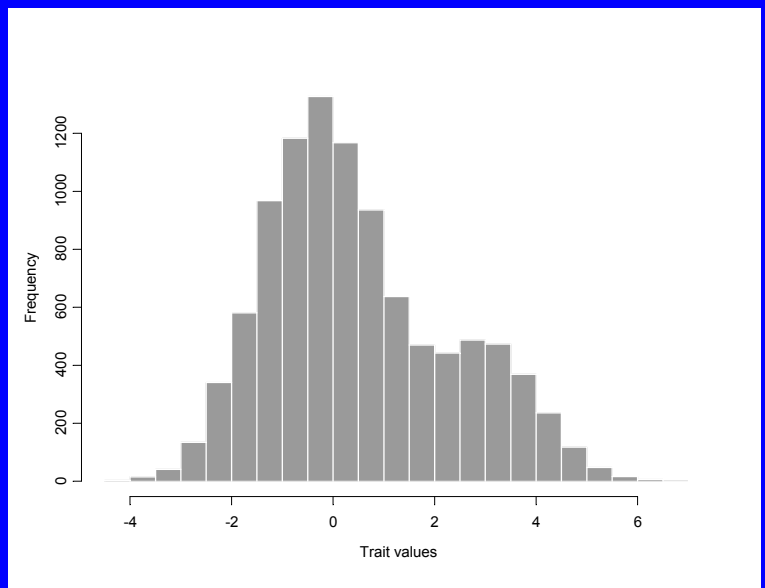
Step 1: Plot the phenotypic data



Get excited! Major gene?



Pity! Many genes of small action



Hint of dominant major gene action? (Pitfall: assumed symmetric/normal traits; may log transform but then lose power – need optimal transformation)

It's a mixture problem…

- Simple model for single QTL with *additive* allele effect and normal error in an F_2.
- Let y_i denote the trait value of the ith individual. We do not know its QTL genotype.
- 3 possible genotypes: $A=a_1a_1$, $H=a_1a_2$, $B=a_2a_2$

$$f(y_i) = 0.25N(\mu_A, \sigma^2) + 0.5N(\mu_H, \sigma^2) + N(\mu_B, \sigma^2)$$

Likelihood $L = \prod_{i=1,..,n} f(y_i)$

Fitting mixtures

- Maximum likelihood: set first-order derivatives of the log likelihood to zero. Find:

$$P(A \mid y_i)\frac{\partial}{\partial\theta}\log(\phi_A(y_i) + P(H \mid y_i)\frac{\partial}{\partial\theta}\log(\phi_H(y_i) + P(B \mid y_i)\frac{\partial}{\partial\theta}\log(\phi_B(y_i)$$

  Problem! Cannot be solved analytically.

- First solution: EM algorithm (Dempster et al 1977) – incomplete data approach, information on QTL genotype is missing (with three components A, H, B)
  - (E-step) Specify or update weights $P(A|y_i)$, $P(H|y_i)$, $P(B|y_i)$.
  - (M-step) Update estimates of μ_A, μ_B, σ^2.

EM Algorithm

- E-step: conditional probabilities calculated using the current parameter estimates
- M=step: weighted least squares

$$\hat{\mu}_A = \frac{\sum_{i=1}^{n}\left[P(A \mid y_i)y_i + P(H \mid y_i)\frac{1}{2}y_i\right]}{\sum_{i=1}^{n}\left[P(A \mid y_i) + P(H \mid y_i)\frac{1}{2}\right]}, etc$$

$$\hat{\sigma}^2 = \frac{1}{n}\left[P(A \mid y_i)(y_i - \hat{\mu}_A)^2 + P(H \mid y_i)(y_i - \frac{1}{2}(\hat{\mu}_A + \hat{\mu}_B)^2 + P(B \mid y_i)(y_i - \hat{\mu}_B)^2\right]$$

- Can do data completion (augmentation) and parameter estimation via iterative reweighted least squares.
- *OR…MCMC (Bayesian or ML)*

Questions

- Want to discover about QTLs underlying quantitative variation for our trait(s) of interest. How many genes are involved? Where are they located on the chromosome? What type of (inter)action do they show?
- Partial information is provided by molecular markers.
- Molecular marker is a locus on the genome where the genotype can be observed with molecular tools: categorical variable with observable state. (QTL is categorical variable with unobserved state.)

Methods of Analysis

- Estimation and testing of means: ANOVA, regression (careful of overparametrisation - more parameters than needed to represent the effects)
- Mixture models via EM or MCMC (Bayesian or frequentist) and choice of model (number of QTL) via single marker interval mapping, composite interval mapping, multiple-QTL mapping
- Problem of multiple testing: too many false positives (non-existing QTLs) – use an experiment-wise error rate, permutation strategies, bootstrap strategies, FDR, re-evaluate the loss function.

Hurray! I've found a QTL!

Statistical association, not a gene. At least four traps:

1. Ghost QTL (error of type 1): two or more linked QTLs with effects of equal sign (QTLs in coupling phase), so not unlikely that the analysis reveals a single QTL in the middle of two true QTLs.
2. Ghost QTL (less anticipated): Unlinked major QTL has inflated the test score. Incidental association can arise due to deviations from expected segregation ratios for any pair of loci on the genome.
3. Nastier (multi-QTL extension): nothing more than an average effect of all QTLs in the region under study, many possibly small QTL effects.
4. Variable information content: if the information content is relatively low in a region containing a QTL, the peak is shifted towards more informative regions.

QTL in outbred pedigrees

- **Outbred or complex pedigrees:**
  - Not formed recently by line crossing
  - Pedigree information available over multiple generations is used instead
  - Eg: milk production in dairy cattle; cholesterol measures in humans
- **Aims of analysis** include linkage mapping, linkage disequilibrium (LD), combined linkage/LD mapping

Comparison with inbred lines

For a moderate-resolution marker map (eg, 10cM):

- disequilibrium measures between QTLs and markers must be expected to be zero.
- distances often exceed 1cM, so any disequilibrium will have eroded over time
- Hence QTL effects cannot be estimated across the population but rather within parents, or phase-known QTL genotypes must be inferred for each parents.

Making inferences

- Consider a parent heterozygous at marker M with alleles M_1, M_2; large number of offspring from this parent (eg,half-sib design in cattle)
- Compare phenotypes of the two offspring groups inheriting the alternative marker alleles (assume biallelic QTL for now):
  - M_1 offspring have higher average phenotype (ie, the allele increasing the phenotype is linked with the M1 allele); similar inference for M_2
  - No detectable difference between the two groups (parent is homozygous at the QTL)

Issues in modelling

- Degree of informativeness of the markers and of the QTL
- Unknown inheritance
- Unknown phases
- Degree of heterozygosity at a QTL
- Multiple families: (i) analyse each family separately; (ii) analyse all families jointly
- Complexity of model depending on the structure of the population

Issues in Modelling

- Power to detect a QTL is limited in outbred pedigrees by the degree of informativeness of the markers and of the QTL (measured by PIC, Botstein et al 1980): heterozygosity of parents combined with fraction of offspring for which the inheritance at a marker is known.
- Inheritance is unknown if an offsprint has the same marker genotype as both of its parents or as one of the parents with the other parent unknown.
- With multiple linked markers, phases will be unknown and need to be inferred.
- Degree of heterozygosity at a QTL is also influential in its detection: if very high or low, a pedigree will contain only families that are not segregating for this allele and hence the QTL will not be detected.
- Multiple families: (i) analyse each family separately; (ii) analyse all families jointly
- Complexity of model depends on the structure of the population: individual large families or small no. families + ignore genetic ties among families, versus multigenerational pedigrees with substantial amounts of missing data

Bayesian approach to linkage mapping

- We want to estimate genotype, recombination rate, etc
- We have
  - unknown quantities U
  - known quantities K

Normalising constant

$$\Pr(U|K) = \Pr(K|U)\ \Pr(U) / \Pr(K)$$

Posterior Likelihood Prior on unknowns

What is known, unknown?

- Unknown quantities U:
  no. QTLs, ordered (phase-known) genotypes of all individuals at all QTLs, ordered genotypes of all individual at all markers; QTL locations modeled as linkage status (located on particular marked chromosome or in residual unmarked chromosome), map positions, map positions of markers, allele frequencies at the QTLs and at markers, QTL genotypic effects, dominance effects, systematic environmental effects, parameters of residual distribution of the phenotypes, parameters of the polygenic background variation.
  (no. alleles at a QTL)
- Known quantities K:
  phenotypes, observed marker genotypes, pedigree

Mapping a monogenic trait

- Vieland (1998): single marker, single trait gene
- Data D=[observed genotypes at marker M and at trait locus T]
- Question: are two loci linked?
  H_0: r=0.5 (no linkage) vs H_L: 0<r<0.5
- δ: genetic distance related to r;
  δ_{min}: min. distance for which r=0.5
- Prior prob. of linkage is Pr(M&T locus on same chromosome) × Pr($\delta<\delta_{min}$)

$$Pr(H_L|D) = Pr(H_L)\ Pr(D|H_L) / P(D)$$

Relation to LOD score

Antilog of LOD score

$$LR = \frac{\Pr(D \mid r = r_{\sup})}{\Pr(D \mid r = 0.5)}$$

$$PostOR = \frac{\Pr(r = r_{\sup} \mid D)}{\Pr(r = 0.5 \mid D)} = \frac{\Pr(r = r_{\sup})}{\Pr(r = 0.5)} \frac{\Pr(D \mid r = r_{\sup})}{\Pr(D \mid r = 0.5)}$$

$$\frac{\Pr(r < 0.5 \mid D)}{\Pr(r = 0.5 \mid D)} = \frac{\Pr(H_L \mid D)}{\Pr(H_0 \mid D)} \frac{\Pr(D \mid H_L)}{\Pr(D \mid H_0)}$$

Priors

- Choose carefully so they do not lead to improper posterior distributions
- Sensitivity analysis: Different priors lead to practically identical inferences if there is sufficient information in the data about the unknowns. If different priors lead to different answers, we need more data.
- Use biologically meaningful prior information in priors.
  Eg, avoid irrelevant frequentist H_0: no QTLs segregating in the entire genome.

Priors for QTL analysis

- Number of QTLs: Poisson
- Linkage indicators of different QTLs independent a priori, = length of marked chromosome / total genetic length of chromosome
- QTL positions, conditional on linkage status, independent a prior, uniform over the length of the marked chromosome
- QTL allele frequencies independent Beta(1,1) (equivalent to Uniform (0,1)).
- Finite polygenic model for residual variation
- etc

Results from a typical run for a single chromosome

- Hoeschele (2001)

-|---|--X--|----X----|---|---|--X--|-----|- 16000
-|---|--X--|----X----|---|---|--X--|--X--|- 2000
-|---|--X--| ------- |---|---|--X--|-----|- 600
-|---|--X--|----X----|---|---| ---- |-----|- 1000
-|---|--X--| ------- |---|---|--X--|--X--|- 400
-|---|-1.0-|--.95 --|---|---|-.95--|-.17--|- intervals
-|---|-----|2:.95,1:.05|--|---|1:.88,2:.12|- regions
Chromosome: 2 QTL: .03; 3 QTL: .87; 4 QTL: .10

Genotype sampling in complex pedigrees

- Genotypes are multilocus, phase-known or ordered, including linked markers and QTLs.
- Want to obtain genotype samples from the joint distribution of genotypes of all pedigree members and at all loci, conditional on observed genotype data (on marker loci) and phenotypic data (y).
- Need these samples in implementation of ML and Bayesian mapping methods.

Peeling

- Genotypic peeling: Conditional probabilities are calculated in a particular order (a 'peeling sequence') for which all conditionals simplify such that each depends on the genotypes of at most two other individuals in the same nuclear family).

$$\Pr(G \mid y) = \Pr(G_1 \mid y)\Pr(G_2 \mid G_1, y_2, .., y_n)\Pr(G_3 \mid G_2, G_1, y_3, .., y_n)...$$

- Also allelic peeling

Fine mapping

- Not often feasible to assign a gene to region of 0.3 cM or less (required for positional cloning, for example) with chromosome dissection methods.
- Use historical recombinations or linkage disequilibrium (LD) (Weir, 1996)
- Different types of LD, influenced by multiple factors (selection, admixture, genetic drift, mutation, migration, coancestry, population expansion etc)

Fine mapping

- Here, marker and phenotypic data are available only on the current generation(s) but there is no pedigree information relating current generation individuals back to ancestral haplotypes carrying a unique, mutant trait allele.
- To date, mostly single marker statistics; some multiple linked markers (Meuwissen&Goddard 2000); some model evolutionary history of population (Lam et al 2000)
- Most applications are for young, rare diseases, not applicable for QTLs.
- Focus on combining LD with linkage mapping.

Why a Bayesian approach?

- Accounts for all uncertainties in the system (eg unknown no. QTLs, unknown genotypes, unknown QTL locations).
- Inferences about particular unknowns of interest obtained conditionally on the observed data but not on particular values of the other unknowns
- Point estimates, marginal posterior distributions, posterior summary statistics etc can be obtained as probability statements.
- Bayesian mean estimator of QTL variance in a marker interval can be interpreted as a multiple shrinkage estimator.