

The University of Newcastle

Kerrie Mengersen

*Introduction to
Bayesian Methods for
QTL Analysis - 5*

(Partial) literature review

- Ball, R. D. (2001). Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. {Genetics} {159}, 1351–1364.
- Bink, M. and Arendok, J. (1999) Detection of quantitative trait loci in outbred populations with incomplete marker data. Genetics.
- George, A. W., Mengersen, K. L., and Davis, G. P. (2000). Localization of a quantitative trait locus via a Bayesian approach. {Biometrics} {56}, 40--51.
- Haley, C. S., and Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. {Heredity} {69}, 315--324.
- Heath, S. C. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. {American Journal of Human Genetics} {61}, 315--324.

- Jansen, R. C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* {135}, 205--211.
- Jansen, R. C., and Stam, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* {136}, 1447--1455.
- Kao, C. H., Zeng, Z.-B., and Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* {152}, 1203--1216.
- Lander, L. S., and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* {81}, 214--224.
- Ma, Casella, Wu (2002) Functional mapping of QTL underlying the character process: a theoretical framework. *Genetics* 161, 1751-1762.
- Mackinnon, M. J., and Weller, J. I. (1995). Methodology and accuracy of estimation of quantitative trait loci parameters in a half-sib design using maximum likelihood. *Genetics* {141}, 755--770.
- Perez-Enciso, M. (2003) Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. *Genetics* 163, 1497-1510.

- Satapogan, J. M., and Yandell, B. S. (1996). Estimating the number of quantitative trait loci via Bayesian model determination. Special Contributed Paper Session on Genetic Analysis of Quantitative Traits and Complex Diseases, Biometric Section, Joint Statistical Meeting, Chicago.
- Satapogan, J. M., Yandell, B. S., Newton, M. A., and Osborn, T. C. (1996). A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. {Genetics} {144}, 805--816.
- Sen, S., and Churchill, G. A. (2001). A statistical framework for quantitative trait mapping. {Genetics} {159} 371--387.
- Sillanpaa, M. J., and Arjas, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. {Genetics} {148}, 1373--1388.
- Sillanpaa, M. J., and Arjas, E. (1999). Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. {Genetics} {151}, 1605--1619.

- Sisson, S. A., and Hurn, M. A. (2003). Bayesian point estimation of quantitative trait loci. Unpublished.
- Stephens, D. A., and Fisch, R. D. (1998). Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. {Biometrics}{54}, 1334--1347.
- Uimari, P., Thaller, G., and Hoeschele, I. (1996). The use of multiple markers in a Bayesian method for mapping quantitative trait loci. {Genetics} {143}, 1831--1842.
- Uimari, P., and Hoeschele, I. (1997). Mapping linked quantitative trait loci using Bayesian method analysis and Markov chain Monte Carlo algorithms. {Genetics} {146}, 735--743.
- Weller, J. I. (1986). Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. {Biometrics} {42}, 627--640.

- Yi, N., and Xu, S. (2000). Bayesian mapping of quantitative trait loci for complex binary traits. {Genetics} {155}, 1391--1403.
- Yi, N., and Xu, S. (2000). Mixed model analysis of quantitative trait loci. {Proceedings of the National Academy of Science, USA} {97}, 14542--14547.
- Yi, N., Xu, S., Allison, D.B. (2003) Bayesian model choice and search strategies for mapping interacting quantitative trait loci. Genetics.
- Zeng, Z. B. (1993). Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. {Proceedings of the National Academy of Science, USA} {90}, 10972--10976.
- Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. {Genetics} {136}, 1457--1468.

Foundation papers

- LANDER and BOTSTEIN (1989): interval mapping
- HALEY and KNOTT (1992) making an approximation to interval mapping using regression, applicable to multiple-QTL models.
- JANSEN (1993), JANSEN and STAM (1994), ZENG (1993, 1994): Composite interval mapping (CIM) and multiple-QTL mapping (MQM)
- KAO et al (1999): multiple interval mapping (MIM).

Bayesian papers

- SATAPOGAN and YANDELL 1996; SATAPOGAN et al 1996; HEATH 1997; UIMARI and HOESCHELE 1997; STEPHENS and FISCH 1998; SILLANPAA and ARJAS 1998, 1999: Bayesian approaches to multiple QTL mapping and reviews.
- GEORGE, MENGERSEN and DAVIS 2000: direct mapping of a QTL
- BINK and ARENDOK 1999: data augmentation for missing data (ungenotyped individuals)
- YI and XU 2000: mapping of a QTL for a complex binary trait

- YI and XU 2000: mixed model analysis of QTLs
- YI, XU and ALLISON 2003: multiple interacting QTL
- YI, GEORGE and ALLISON 2003: stochastic search variable selection for identifying multiple QTL
- SEN and CHURCHILL 2001: a general statistical framework for quantitative trait mapping
- BALL 2001 the use of the Bayesian information criterion in model selection for Bayesian methods of quantitative trait loci mapping
- SISSON and HURN 2003: Bayesian point estimation of QTL
- PEREZ-ENCISO 2003: Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information
- MA, CASELLA and WU 2003: mapping traits that change as a function of some independent variable.

Uimari and Hoeschele

- Bayesian method for mapping linked QTL using multiple linked genetic markers.
- Uses MCMC.
- Parameters included allele frequencies and substitution effects for two biallelic QTL, map positions of the QTL and markers, allele frequencies of the markers, and polygenic and residual variances.
- Missing data were polygenic effects and multi-locus marker-QTL genotypes.
- Three different MCMC schemes for testing presence of a single or two linked QTL on the chromosome. All methods performed well.

Adapted from summaries produced by Emily Anderson (UNcle)

George, Mengersen, Davis

- Direct mapping of a QTL, with full use of information from multiple linked gene markers.
- Estimation of QTL genotype probabilities for sires and offspring; allele frequencies for the QTL; the position and additive and dominance effects of the QTL.
- The ability of the technique used in the paper to estimate the parameters accurately is examined for a variety of scenarios.

Bink and Arendonk

- Augmentation of marker genotypes for ungenotyped individuals, implemented in a Bayesian approach via MCMC.
- Marker data on relatives and phenotypes are combined to compute conditional posterior probabilities for marker genotypes of ungenotyped individuals to detect segregating QTL.
- Allelic effects at the QTL assumed normal, with covariance matrix based on known QTL position and IBD probabilities derived from flanking markers.
- Tested on complex granddaughter design: significant increase in power when ungenotyped dams were included in the analysis.

Yi and Xu

- Focus on complex binary traits, aiming to demonstrate that Bayesian methods are helpful in the mapping of QTLs for complex binary traits.

$$y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \sum_{j=1}^l \mathbf{Z}_{ij}^T \mathbf{H}_j \gamma_j + \varepsilon_i$$

- Linear model:
- Initial modeling of a complex binary trait using the typical threshold model.
- Use Bayesian methods to find the liability (hypothesized variable which underlies the phenotype), and a threshold.
- Posterior samples of the unknowns produced using RJMCMC.
- Estimate joint posterior distribution of number of QTLs, the localization, and the effect of the identified QTLs.

$\boldsymbol{\beta}$ covariates, \mathbf{X} known incidence matrix,

\mathbf{Z} indicators for 4 possible ordered genotypes; $Z_{ijk}=1$ if k th genotype is observed

\mathbf{H} matrix of linear contrasts converting 3 genetic effects into genotypic values of 4 genotypes

MCMC algorithm

1. Update the liability \mathbf{Y} individual by individual
2. Update coefficients β and $\{\gamma_i\}_{i=1}^l$
3. Update the number of QTLs, l , and their locations, λ
4. Update the QTL genotypes individual by individual and locus by locus
5. Update marker genotypes individual by individual and locus by locus

In Step 3 a random choice is made from: modifying the QTL locations, adding a new QTL to the model, and deleting a QTL from the model. The order of the QTLs is not fixed when updating their locations, the components of λ are modified one at a time using the Metropolis-Hastings algorithm. To add a QTL, a new location is sampled, then new genetic effects and QTL genotypes for every individual are sampled. To delete a QTL, one of the existing QTLs are randomly chosen and the corresponding acceptance probability is found.

Comments

- Need to be careful in the interpretation of categorical data with a threshold model since the liability is a hypothesised variable.
- Caution about choice of proposal distribution for the QTL effect: it strongly affects mixing when a new QTL has been added to the model.
- Issue of determining effective sampling sizes in MCMC and assessment of convergence: difficult to ascertain serial correlation due to changing dimension of each cycle.
- In general, the single-site updating (updating the genotype individual by individual and locus by locus) does not always lead to an irreducible sample, due to the strong dependency of adjacent loci.
- Epistatic effects not included, but conceptually easy to do so.

Yi and Xu

- Mixed model analysis for mapping of QTLs for a hybrid population derived from two or more distinct outbred populations being crossed.
- The mean allelic value of each of the source populations is treated as the fixed effect and the allelic deviations from the mean are treated as the random effects in the mixed model approach.
- Allows partition of the total genetic variance into between-population variance and within-population variance.
- Bayesian techniques are used in statistical inference of the QTL parameters via Markov chain Monte Carlo.

Yi, Xu and Allison

- Use Bayesian model and variable selection to develop strategies for identifying multiple QTL with complex epistatic patterns in experimental designs with two segregating genotypes.
- RJMCMC to determine number of QTL and to select main and epistatic effects.
- Method can map a large number of QTL with any combination of main and epistatic effects.
- Sensitivity of posterior inference to prior specifications of the number and genetic effects of QTL is investigated

Yi, George and Allison

- Stochastic search variable selection methodology for identifying QTL for complex traits in experimental designs.
- Embed multiple regression in a hierarchical normal mixture model, where latent indicators for all markers are used to identify the multiple markers. The markers with significant effects can be identified as those with higher posterior probability included in the model.
- Simple Gibbs sampler employed.
- Results show that the method works well under typical situations of most QTL studies in terms of number of markers and marker density.

Sen and Churchill

- General framework for statistically analysing quantitative trait data in inbred line crosses.
- By conditioning on the unknown QTL genotypes, the framework is based on splitting the analysis apart into two sections
 - the relationship between the QTL and the phenotype
 - the location of the QTL in the genome.
- A basic Monte Carlo algorithm is presented to apply the Bayesian analysis used.
- To obtain information in the phenotype data, weights are given to genotypes simulated in the Monte Carlo algorithm.

- Factorise the joint distribution:

the quantitative trait measurements }
the genetic model parameters

the marker data }
the QTL locations

Solve independently,
given

the QTL genotypes

- Posterior distribution of the QTL genotypes is obtained:
after integrating out the genetic parameters and QTL
locations, express it as the product of two terms:

- Compatibility between a phenotype and the QTL genotypes $p(y|g)$
- Compatibility between the QTL genotypes and the known marker data $p(g|m)$.

So: sample from $p(g|m)$, weight by $p(y|g)$

Ball

- An approximate method for QTL mapping analysis.
- The methods basis is model selection from multiple regression models with trait values regressed on marker genotypes.
- This uses a modified Bayesian information criterion to approximate the posterior probability of models using a range of subsets of markers as variables. The BIC- δ criterion is also modified to include prior information; missing values dealt with using multiple imputation.
- Paper gives marginal probabilities for the different model sizes
- Marginal probability of a QTL being in a region is estimated by the probability that one or more markers in that region are selected: found using the BIC by summing the posterior probabilities for models containing one or more of the markers.

Sisson and Hurn

- Paper focuses on the locations of positions of the best candidate markers segregating for the trait, and introduces a loss function for estimating the number of QTLs and their locations.
- Think of the (countable) union of spaces corresponding to no QTL, one QTL etc. Let ϕ , ϕ^* be two such spaces. The loss function $L(\phi, \phi^*)$ is the loss or error made when estimating ϕ using ϕ^* .
- Bayes estimates by definition minimise the expected posterior loss. Difficult over varying dimensionality.
- Sisson and Hurn argue that commonly used loss functions are unsuitable for estimating QTL position because the loss function:
 - (i) should hold QTL locations in greater importance than the order in which the points occur, and
 - (ii) should be able to handle varying dimensionality.

- Adapt a loss function proposed by Celeux, Hurn and Robert (2000) for estimation of parameters of unlabelled mixture distributions.
- Begin by defining a large number of “control points” t_1, \dots, t_T , belonging to the same space as the components of the mixture. The loss function is then given as

$$L(\phi, \hat{\phi}) = \sum_{i=1}^T [d(t_i, \phi) - d(t_i, \hat{\phi})]^2$$

where $d(t_i, \phi)$ is the Euclidean distance between the control point t_i and the nearest of the components of ϕ . If the distance from any t_i to the nearest ϕ_j component is not equal to its distance to the nearest $\hat{\phi}_j$ component, this form of the loss function gives input to the loss.

- Adapt this for a single chromosome scenario:

$$d'(t_i, \phi) = \begin{cases} d(t_i, \phi) & \text{if } |\phi| \geq 1 \\ d(t_i, t_i + K) & \text{if } |\phi| = 0 \end{cases}$$

where $|\phi|$ is the number of loci in configuration ϕ and K is the length of the chromosome. Then the loss function is given by

$$L(\phi, \tilde{\phi}) = \sum_{i=1}^T [d'(t_i, \phi) - d'(t_i, \tilde{\phi})]^2.$$

So here the loss function accounts for difference between $|\phi|$ and $|\tilde{\phi}|$ by the contribution of those t_i whose nearest QTL location changes between the two configurations.

The paper then considers the scenario of estimating an unknown number of QTLs located over C chromosomes. A "between-chromosome spatial structure" on the genome is proposed

Perez-Enciso

- Bayesian method that combines linkage and linkage disequilibrium (LD) information.
- Method uses jointly all marker information (haplotypes) and all available pedigree information, ie not restricted to any specific experimental design and known phases not required.
- A diallelic QTL is assumed and both additive and dominant effects can be estimated.
- Also implemented a Bayesian variant of the usual disequilibrium measures like D' and r^2 between QTL and markers.
- Using LD information resulted in much better estimates of QTL position when there was complete disequilibrium between mutant QTL allele and the marker; advantage decreased when the association was only partial.

Your turn!