VISG Virtual Institute of Statistical Genetics

Peeling and Bayesian QTL mapping for allo-polyploids

Roderick D. Ball, Scion† and Gail Timmerman-Vaughan, Plant and Food‡

SUMMARY: We develop a general multi-locus Bayesian QTL mapping method for allo-polyploids. Our algorithm uses peeling (Elston and Stewart 1971; the original Bayesian graphical modelling method for evaluating probability distributions on pedigrees) to sample from the posterior distribution of fully informative 'virtual markers'. The virtual markers are then used as input to a Bayesian QTL mapping method. (the BIC method; Ball Genetics 2001; R package `BayesQTLBIC`; `http://cran.r-project.org/web/packages/BayesQTLBIC/index.html`), a non-MCMC method based on using the BIC criterion to obtain good approximations to posterior probabilities for linear models representing possible QTL configurations, with multiple peeling samples analysed jointly using multiple imputation. Single marker peeling is used together with an incremental HMM peeling strategy to sample from the posterior distribution of virtual marker genotypes. By eliminating genotype errors and eliminating low probability combinations at each stage we ensure the computation is quadratic in the number of loci per chromosome and linear in the number of progeny. The method is applied to allo-tetraploid clover data. KEYWORDS: Polyploids, QTL mapping, Bayesian model selection, HMM peeling, multiple imputation.

Introduction

Allo-polyploids are species with more than 2 paired sets of (homoeologous) chromosomes, e.g., a hexaploid, of the form AABBCC, has arisen from ancestral species A, B, C. Many important species e.g. tetraploid clover, hexaploid kiwifruit, redwoods, arabica coffee, cassava, wheat, oats, and salmon are allo-polyploid. In allo-polyploids the sub-genomes (e.g. A, B, C) retain their identity in meiosis (preferential pairing, disomic inheritance). Offspring inherit one copy of ABC from each parent. The goal of quantitative trait loci (QTL) mapping is to find loci associated with traits, given DNA marker data and trait data for a pedigree. QTL mapping is challenging in polyploids because:

- few markers are fully informative (cf. Fig. 2) creating a missing data problem. For example in Figure 2, P1 C1 has a large segment of missing data ('0') due to lack of polymorphic markers, and,
- there are many possible parental combinations of alleles.

Previous (frequentist) approaches to QTL mapping in polyploids are based on single markers or pairs of 'flanking markers (Doerge and Craig 2000; Luo et al. 2000; Hackett et al. 2001; Cao et al. 2004, 2005). Frequentist QTL mapping loses information when flanking markers are not fully informative, and is limited to plotting likelihood ratios along the genome. Our goal is to develop a Bayesian method for allo-polyploids QTL mapping which incorporates the information in all markers jointly and enables calculation of posterior distributions for the number, location, and size of QTL effects.

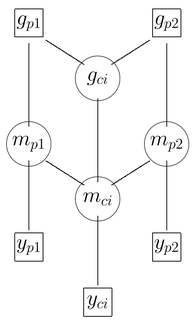
Our approach

- Use a modification of the peeling algorithm (Elston and Stewart 1971) to sample from the posterior distribution of sets of fully informative 'virtual markers' (unobserved fully informative marker genotypes, g_{ci}, consistent with the observed data, Figure 1)
- Use the virtual markers as input to the BIC method in the R package BayesQTLBIC (Ball 2001, 2009).

Peeling–background

- Algorithm for evaluating a probability distribution (Elston and Stewart 1971)
- Based on a graph representing a pedigree, genotypes on nodes (more generally any data)
- Sum over the values of a terminal node, represent the distribution on the remaining nodes
- Reduce to the distribution on single node
- Reverse the process to generate samples
- The computational complexity is exponential in number of alleles $\times$ markers (but linear in the number of individuals)
- Generally computationally feasible for up to 5 markers (in diploids)
- Therefore we need to adapt the peeling algorithm to the allopolyploid situation.

Allo-polyploids peeling hierarchical model



g_{pi}, g_{ci} parental and progeny virtual marker genotypes

m_{pi}, m_{ci} parental and progeny marker genotypes

y_{pi}, y_{ci} parental and progeny marker phenotypes

Figure 1: Hierarchical model for polyploids peeling.

Peeling (cf. Fig 1.)

1. Write down the joint distribution:

$$f(m_{p1}, m_{p2}, \ldots) = [m_{p1}][m_{p2}][y_{p1}|m_{p1}][y_{p2}|m_{p2}]\times \prod_{i=1}^{n}[g_{ci}|g_{p1}, g_{p2}][m_{ci}|g_{ci}, m_{p1}, m_{p2}][y_{ci}|m_{ci}] \quad (1)$$

2. Peel m_{ci}: for each progeny (i), sum over m_{ci}, represent the sum of relevant terms as a function $R_{gci}(g_{ci}, m_{p1}, m_{p2})$, giving an equation for the marginal distribution $f(m_{p1}, m_{p2}, g_{ci})$.
3. Peel g_{ci}: for each progeny (i), sum over g_{ci}, represent the sum of relevant terms as a function $R_{gp1p2i}(g_{p1}, g_{p2}, m_{p1}, m_{p2})$ giving the marginal distribution $f(m_{p1}, m_{p2})$ of parental genotypes.

Apply for single loci (single marker peeling), or multiple loci (use HMM peeling for each progeny), cf. algorithm overview.

Algorithm overview

Single marker peeling

- Peel individual markers separately
  - + genotype errors can be identified by fitting impossible genotypes with very low probability. If impossible genotypes occur, many low probability parental combinations will occur in the output. In contrast, with no genotype errors, there will be a modest number (typically up to 8 for allo-tetraploids) of combinations with non-negligible probability for each parent.
- Restrict to parental combinations with non-negligible probability

↓

HMM peeling

- Conditional on parental genotype combinations, individual progeny genotypes for a chromosome are an HMM
- HMM computation $O(N^2T)$ for T loci, ploidy $2k$ implies $N = 2^{2k}$ states per locus
- But, the number of parental allelic combinations is exponential
- Therefore use an iterative strategy, $O(N^2T^2)$ per progeny:
  - + ensure that the number of parental combinations selected is bounded at each step. This is typically 64 (for 1 marker phase alignment class) or 128 (if there are 2 alignment classes) in allo-tetraploid clover.
  1. Let $t \leftarrow 2, \texttt{min\_prob} > 0$, e.g. $\texttt{min\_prob} = 5 \times 10^{-4}$.
  2. For $p \in \texttt{parental\_combinations}(1, \ldots, t-1) \times \texttt{parental\_combinations}(t)$
  3. do HMM for loci $1 \ldots t$ for each progeny conditional on p
  4. calculate $\Pr(p) =$ marginal probability from HMM
  5. eliminate low probability combinations ($\Pr(p) < \texttt{min\_prob}$).
  6. Increment t and repeat steps 2–5 until $t = T$
  7. Sample parental combinations
  8. Repeat HMM to sample progeny virtual marker genotypes (g_{ci} in Fig. 1)
  9. Align marker phases on samples, group into alignment classes

↓

BIC method for QTL mapping

- R package `BayesQTLBIC` (Ball 2001, 2009).
- Poisson prior, $n_{QTL} \sim$ Poisson(λ_Q), e.g. $\lambda_Q = 10$
- Prior probability per marker is proportional to marker spacing, n_{QTL}
- Model matrix X is populated with contrasts from virtual marker genotypes
- Models $\mathcal{M}_\gamma : y = X_\gamma\beta_\gamma + \epsilon, \quad X_\gamma \leftrightarrow$ selected columns of $X \leftrightarrow$ putative QTL
- Exhaustive search through model space (option to control model size, e.g. ≤ 3 QTL per chromosome) compute: $\Pr(\mathcal{M}_\gamma) \propto \exp(-\text{BIC}/2) \times$ prior, for each model
- Missing marker values are handled by multiple imputation, BIC is adjusted for the number of imputations
- The multiple imputations correspond to multiple samples of virtual marker genotypes (g_{ci}) from HMM peeling,
- Compute probabilities e.g.

$$\Pr(\text{QTL in region}) = \sum\{\Pr(\mathcal{M}_\gamma) \text{ with QTL in region}\} \quad (2)$$

- Sample from models for each chromosome and recalculate probabilities to obtain a combined model for whole genome (optional).

Application to allo-tetraploid white clover data

- White clover data (Barrett et al. 2004, 2005) allo-tetraploid
  - + n=182 progeny, m=152 markers, 8 chromosomes, map length 592cM
  - + traits analysed here: seed yield 2002, 2003, 2004; mean seed yield
  - + traits measured in replicated field trials at Lincoln, New Zealand
- Peeling: parental genotypes for C1C2, sample 1 shown in Fig. 2
- Good evidence for QTL on chromosomes C, D, year dependent (Tables 1,2)

Table 1: Bayes factors[1] for seed yield, chromosomes C and D.

	year	C1		C2		D1		D2	
seed yield	2002	0.56	0.04	0.03	164892	0.50	0.18	10.0	0.05
	2003	0.53	0.03	0.06	0.16	0.43	0.07	63207	0.03
	2004	0.70	0.30	0.03	0.06	0.18	0.09	966859	0.06
	mean	0.70	0.04	0.03	11.96	0.83	0.16	7195	0.04

[1] Bayes factors testing for 1 or more QTL on each chromosome, separate analyses per parent per subgenome.

Table 2: Marginal probabilities for model sizes for mean seed yield.

	model size						model size				
	0*	1	2	3	4		0	1	2	3	4
A1A2	1.1×10^{-1}	0.54	0.32	0.03	< 0.01	E1E2	4.8×10^{-1}	0.39	0.12	0.01	< 0.01
B1B2	3.2×10^{-1}	0.53	0.14	< 0.01	0.01	F1F2	5.6×10^{-1}	0.36	0.08	< 0.01	< 0.01
C1C2	1.6×10^{-3}	0.04	0.25	0.66	0.05	G1G2	3.7×10^{-2}	0.66	0.27	0.03	< 0.01
D1D2	1.1×10^{-5}	0.26	0.46	0.27	< 0.01	H1H2	1.0×10^{-1}	0.66	0.21	0.02	< 0.01

*Cf. prior probability of H_0 per chromosome $\pi_{H_0} \in (0.003, 0.015)$.

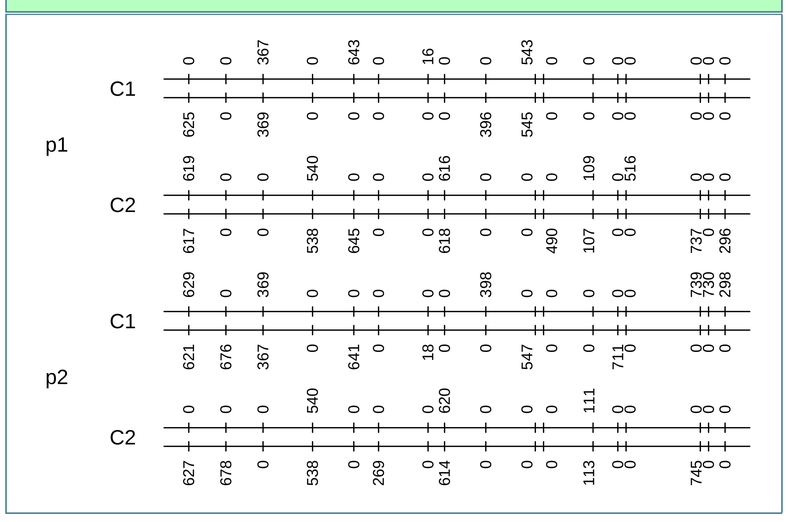


Figure 2: Parental (p1,p2) genotype combinations from peeling for chromosomes C1 and C2, sample 1.

Conclusions

By combining a modified peeling algorithm with the BIC method we have developed the first general multi-locus Bayesian method for QTL mapping in allo-polyploids with the following benefits:

- Uses information from all markers jointly
- Can detect genotype errors
- Can compute Bayes factors and posterior probabilities for number of QTL per region or chromosome
  - + therefore avoid spurious associations due to significant p-values in frequentist methods when Bayes factors or posterior probabilities not high
- Provides model averaged QTL effects and effects of allelic substitution
  - + therefore avoid selection bias ('Beavis effect') common with frequentist QTL mapping methods (cf. Beavis 1994; Ball 2001)
- Is computationally feasible giving approximate probabilities using HMM peeling by successively eliminating low probability parental combinations
  - + peeling: 6–400 seconds per chromosome for tetraploid clover
- Computational complexity $O(N^2T^2), N = 2^{2k}$ for ploidy $2k$, T loci per chromosome,
- Can in principle extend to other types of polyploidy, by changing the segregation model.

References

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.

BALL, R. D. 2009: BayesQTLBIC—Bayesian multi-locus QTL analysis based on the BIC criterion. `http://cran.r-project.org/web/packages/BayesQTLBIC/index.html`

BARRETT, B., GRIFFITHS, A., ... and WOODFIELD, D. 2004: A microsatellite map of white clover. Theor. Appl. Genet. 109: 596–608.

BARRET, B. A., BAIRD, I. J., and WOODFIELD, D. R. 2005: A QTL analysis of white clover seed production, Crop Sci. 45: 1844–1850.

BEAVIS, W. D., 1994: The power and deceit of QTL experiments: lessons from comparative QTL studies. Proc. 49th Ann. Corn and Sorghum Indus. Res. Conf.

CAO, D., CRAIG, B. A., and DOERGE, R. W. 2005: A model selection based interval mapping method for autopolyploids. Genetics 169: 2371–2382.

DOERGE, R. W. and CRAIG, B. A. 2000: Model selection for quantitative trait locus analysis in polyploids. PNAS 14: 7951–7956.

ELSTON R. C. and STEWART, J. 1971: A general model for the genetic analysis of pedigree data. Human Heredity 21: 523–542.

HACKETT, C. A., BRADSHAW, J. E. and MCNICOL, J. W. 2001: Interval mapping of quantitative trait loci in autotetraploid species, Genetics 159: 1819–1832.

LUO, Z.W., HACKETT, C.A., BRADSHAW, J.E., MCNICOL, J.W. and D.MILBOURNE, D. 2000: Predicting parental genotypes and gene segregation for tetrasomic inheritance, Theor. Appl. Genet. 100: 1067–1073.

Acknowledgements

This work was funded by the New Zealand Foundation for Research Science and Technology through a contract with the Virtual Institute for Statistical Genetics (VISG).

For further information contact: R. D. Ball, Scion,† P. B. 3020, Rotorua, New Zealand. Email: `rod.ball@scionresearch.com`. URL: `www.scionresearch.com`.

†Scion is a trading name of the New Zealand Forest Research Institute Limited.

‡Plant and Food is a trading name of the New Zealand Plant and Food Research Institute Limited.

Presented at BayesComp, Tokyo; and ISBA, Kyoto June 2012.