A strategy for multiple linkage disequilibrium mapping methods to validate additive QTL

Li Yi[1,4], Jong-Joo Kim[2], Kwan-Suk Kim[3]

[1]School of Statistics, Shanxi University of Finance & Economics, Taiyuan, Shanxi, China
[2]School of Biotechnology, Yeungnam University, Gyeongsan, Gyeongbuk, Korea
[3]Department of Animal Science,Chungbuk National University, Cheongju, Korea
[4]Corresponding author: Li Yi, e-mail: liyi3344520@hotmail.com

Abstracts

The efficiency of genome-wide association analysis (GWAS) depends on power of detection for quantitative trait loci (QTL) and precision for QTL mapping. In this study, three different strategies for GWAS were applied to detect QTL for carcass quality traits in the Korean cattle, Hanwoo; a linkage disequilibrium single locus regression (LDRM), a combined linkage and linkage disequilibrium analysis (LDVCM) and a BayesCπ approach. Phenotypes of 486 steers were collected for carcass weight, backfat thickness, *longissimus dorsi* muscle area, and marbling score, and genotype data were also scored with the Illumina bovine 50k SNP chips for the steers. For the former two methods, threshold values were set FDR <0.01 on chromosome-wise level , while a cut-off such that top 5 variance of 10-SNP windows that were included in the Bayes Cπ model were determined for the latter model. A total of 12 and 10 QTL were detected by LDRM and LDVCM respectively, and after BayesCπ refining four consistent QTL were obtained, which corresponded to QTL regions previously reported. Some well-known candidate genes for the traits of interest were located close to these QTL. Our result suggests the use of combined different LD mapping approaches can provide more reliable chromosome regions to further pinpoint DNA makers or causative genes in these regions.

Key words: QTL, Linkage Disequilibrium, SNP, Hanwoo

1.Introduction

The availability of genome-wide SNP panels enables detection of statistical associations between a trait and any SNP in terms of a genome-wide association study (GWAS), enhancing the possibility of mapping QTL across the genome. In the literature, several statistical methods have been suggested and applied (Bayesian vs. frequentist methods) (Sun et al.2011;Zhao et al.2007; Cierco-Ayrolles et al.2010; Erbe et al.2011;Uleberg et al.2010). These methods are based on assumption that saturated marker have increased the feasibility of QTL detection and mapping using historical population-wide linkage disequilibrium, which requires a marker allele to be in LD with the QTL allele across the entire population. The term ‘linkage disequilibrium’ seems to imply disequilibrium of two loci being in physical linkage, but disequilibrium can also exist between two physically unlinked loci. Such associations may lead to collinearity among these SNPs and may impair reliability of LD-based QTL mapping as they may generate significant signals far from the real QTL position. With such a large number of tests being performed, we expected a large number of false positives.

The aim of this study was to minimize the chance of collinearity and overcome the tendency to find significant signals from a causal variant by using the strategy for multiple linkage disequilibrium mapping methods to validate additive QTL. With this approach, Significant results were set FDR <0.01 on chromosome-wise level for the marker-trait association (LDRM) to be regarded as significant. Furthermore, we can use information from linkage to filter spurious associations in the GWAS and further refine intervals containing QTL by using combined linkage disequilibrium and linkage (LDVCM). Finally, we demonstrate this by refining the QTL position for predicting the effects of the SNPs which fits all the SNPs simultaneously (BayesCπ).

2.Methods

Three complementary approaches were used: (i) linkage disequilibrium single locus analysis using the snp_a option of Qxpak (5.03 version) software(Perez-Enciso and Misztal 2011); (ii) combined linkage disequilibrium and linkage analysis using LDVCM; (iii) Bayesian analysis using the BayesCπ option of GenSel (http://bigs.ansci.iastate.edu/bigsgui/login.html).

LDRM First, linkage disequilibrium single locus regression (Zhao et al.2007;Grapes et al.2004)were performed.

$$y = \mu + X\alpha + Z_u u + e$$

Where y is the phenotypic record, μ is the average phenotypic performance, X is the design matrix of SNP genotype (e.g. individuals with marker genotypes '11', '12' and '22' are assumed to have genetic values μ_{kAA} , 0 and μ_{kBB}), a is the fixed substitution effect for the SNP, Z_u is the incidence matrix for animal effects, u is the infinitesimal genetic effect, which is distributed as $N(0,A\sigma_u^2)$ (the numerator relationship matrix A and the additive genetic variance σ_u^2) and e is a random residual for animal *i*, which is distributed as $N(0, I\sigma_e^2)$ (the identity matrix *I* and residual variance σ_e^2). Likelihood ratio tests were performed by removing the single locus SNP genotypic effects, and P-values were obtained assuming a χ^2 distribution of the likelihood ratio test with one degree of freedom. Association with false discovery rate <0.01 on chromosome-wise level were considered significant.

LDLA The LDVCM software has been essentially described (Kim and Georges 2002; Blott et al.2003). Briefly, the linkage phases of all sires and sons were determined by the approach of Druet and Georges (2010). Then, identity-by-descent(IBD) probabilities(φ_p) at the midpoint of each SNP interval (*p*) were computed for all pairs of haplotypes conditional on the identity-by-state status of flanking markers (Meuwissen and Goddard, 2001). A dendrogram was generated by using the unweighted pair group method with arithmetic mean (UPGMA) hierarchical clustering algorithm with 1-φ_p as the distance measure at QTL location (*p*).Starting at the ancestral node and sequentially descending into the dendrogram, all possible combinations of haplotype clusters were analyzed in place of individual haplotypes. This process identified the set of nodes at which the likelihood of the data were

maximized.

To jointly exploit linkage disequilibrium (female meiosis) and linkage (male meiosis) information, the following mixed linear was used:

$$y = \mu + Z_h h + Z_u u + e$$

Where *h* is the vector of random QTL effects corresponding to the defined haplotype clusters. Z_h is an incidence matrix relating maternal haplotypes of sons and sire haplotypes to individual sons. Likelihood ratio tests were performed by removing the haplotype cluster effects, and P-values were obtained assuming a χ^2 distribution of the likelihood ratio test with one degree of freedom. Association with FDR <0.01on chromosome-wise level were considered significant.

BayesCπ BayesCπ were developed from the BayesB and GBLUP approaches (Meuwissen et al.2001) and was described in detail by Habier et al.(2011).

$$y = \mu + \sum_{j=1}^{K} X_j \alpha_j \delta_j + e$$

Where K is the number of SNPs, X_j is the vector of genotypes at SNP_j, a_j is the random substitution effect for the SNP_j, which condition on the variance $\sigma^2_{\alpha_j}$, is assumed normally distributed $N(0, \sigma^2_{\alpha_j})$ when δ_j=1, while $\sigma^2_{\alpha_j}$=0, when δ_j=0, $\sigma^2_{\alpha_j}$ is a random 0/1 variable indicating the absence (with probability π) or presence (with probability 1-π) of SNP_j in the model. The prior for π was treated as unknown with uniform (0, 1).Gibbs sampling was applied to calculate the posterior means of model parameters μ, a_j, $\sigma^2_{\alpha_j}$, σ^2_e, and π. The MCMC algorithms were run for 50,000 samples, with the first 20,000 samples discarded as burn in. A window size of 10 was used and the variance of each 10-SNP window was used as the criterion to detect QTL. Several windows that shared the same SNP with a large effect were considered to identify the same QTL region. Within each region, because windows were overlapping, the window with the highest variance of GEBV was used and the SNP within this window that explained the largest proportion of genetic variance was used to denote the position of the QTL. To select which positions should be major QTL, we first ranked them by estimated variance. We then chose a cut-off such that top5 QTL were detected.

Information about particular genes, located near SNP significantly associated with each trait, was extracted from online sources (http://www.ensembl.org/index.html, http://www.genecards.org/cgi-bin/cardsearch.pl#top, and http://www.uniprot.org).

3. Application to the carcass quality in Hanwoo

We applied this strategy to the carcass quality in Hanwoo (486 steers, Six traits and 39,506 SNPs).The 17 QTL detected by LDRM or LDVCM were integrated using the BayesCπ to remove possible redundancy among those QTL. Four QTL from these three methods had high concordance including 64.1-64.9Mb for BTA7 for WWT, 24.3-25.4Mb for BTA14 for CWT, 0.5-1.5Mb for BTA6 for BFT and 26.3-33.4Mb for BTA29 for BFT (Figure 1). The focus of our study was the detection of major QTL position rather than the precise estimation of their

effects. Previous work has shown that these three methods belong to three distinct analytical, *i.e.* LDRM test the single marker at a time and regards the markers as independent of all other markers (Zhao et al.2007; Grapes et al.2004; MacLeod et al.2010), LDVCM test the mid-point of the marker brackets, which corresponded best when the QTL was masked between analyzed markers (Kim and Georges 2002; Blott et al.2003), and BayesCπ test the effects of all markers are fitted simultaneously. However, QTL with large effects can be detected by both the bayesian shrinkage and linear regression mapping methods. By specifying proper prior distributions for SNP effects, the ignorable small SNP effects are coerced to zero and only SNPs with larger effects on phenotype are fitted in the model, hence Bayesian shrinkage analysis could reduce possible spurious QTL effects by adjusting all other QTL effects. This was also explained by Xu (2003) and Sun et al. (2011). Therefore, here we attempted to minimize the number of false positives by combined these three methods result.

The establishment of a threshold is a complicated issue and has a profound influence on the results, especially when methods with different nature of scores are compared. False discovery rate derived threshold takes into account the number of tests that are performed as well as how significant one test is relative to the others in multiple comparison procedures. Two levels of significant controls were used in LDRM and LDVCM analysis based on genome-wise (FDR<0.05) and chromosome-wise (FDR<0.01) type I errors, which were computed for all SNP(Fernando et al.2004). Usually, significance tests are not required in Bayesian analysis; only frequentists emphasize significance tests. However, to facilitate comparison with other methods, we accumulated the effects of adjacent SNPs together into a genomic window and then chose a cut-off such that top 5 as the criterion to detect QTL in BayesCπ. Figure 1 clearly shows the major peaks in BayesCπ coincide with LDRM and LDVCM. Therefore, These present findings together with those of Sun et al.(2011) suggest that the genetic variances estimated by BayesCπ including all markers without polygenic effects can exploit the gene discovery knowledge generated.

Our Study uncovered successfully many variants related to QTLs. However, the traits YWT, CWT and BFT measured in Hanwoo have QTL with larger additive effects than WWT, LMA and Marb. Several explanations have been proposed for our scant outcomes from genetic markers. First, the currently identified SNPs might not fully describe genetic diversity. For instance, these SNPs may not capture some forms of genetic variability that are due to copy number variation. Second, genetic mechanisms might involve complex interactions among genes and between genes and environmental conditions, or epigenetic mechanisms which are not fully captured by additive models. However, opportunities may exist for improving predictions by exploiting additive genetic variation. A third explanation — the one we focus on here — lies in the limitations posed by the genetic models and statistical methods.

4. Conclusions

In this work, no linkage disequilibrium mapping method tested in this study outperformed the others, but it is interesting to that this position of the SNP selected by the different models (LDRM, LDVCM, BayesCπ) were close. Some well-known candidate genes for the traits of interest were located close to these QTL. We suggest the use of combined different LD mapping approaches can provide more reliable chromosome regions to further pinpoint DNA makers or causative genes in these regions.

5. Acknowledgements

This research was supported by Shanxi Scholarship Council of China（2013-072）.

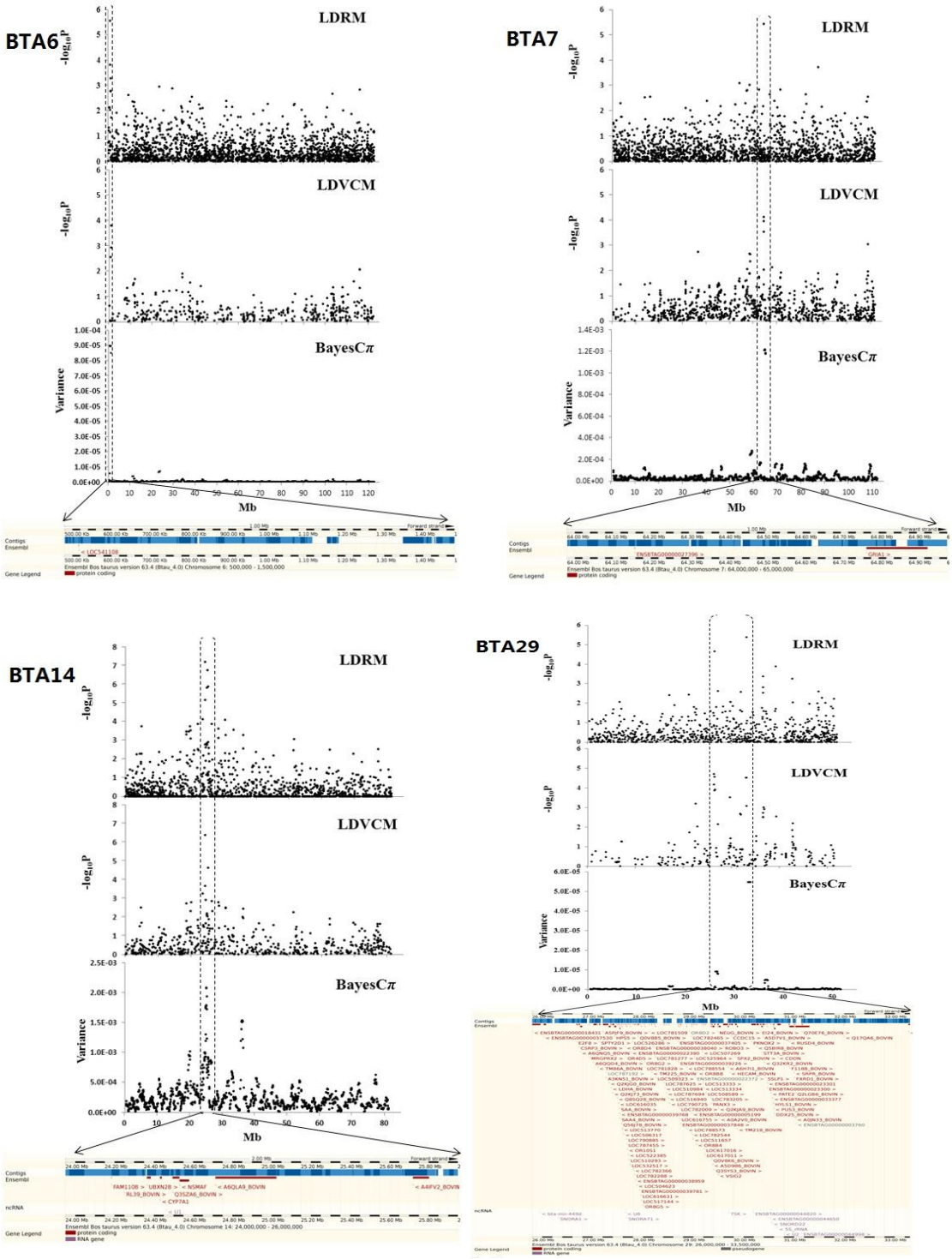

Figure 1.QTL profile (upper) and annotation (lower) of the detected SNPs associated with BFT on BTA6, WWT on BTA7, CWT on BTA14 and BFT on BTA29

References

Blott, S., et al. (2003), "Molecular Dissection of a Quantitative Trait Locus: A Phenylalanine-to-Tyrosine Substitution in the Transmembrane Domain of the Bovine Growth Hormone Receptor Is Associated with a Major Effect on Milk Yield and Composition," Genetics, 163, 253-266.

Cierco-Ayrolles, C., et al. (2010), "Does Probabilistic Modelling of Linkage Disequilibrium Evolution Improve the Accuracy of Qtl Location in Animal Pedigree?," Genet Sel Evol, 42, 38.

Druet, T., and Georges, M. (2010), "A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping," Genetics, 184, 789-798.

Erbe, M., Ytournel, F., Pimentel, E. C., Sharifi, A. R., and Simianer, H. (2011), "Power and Robustness of Three Whole Genome Association Mapping Approaches in Selected Populations," J Anim Breed Genet, 128, 3-14.

Fernando, R. L., et al. (2004), "Controlling the Proportion of False Positives in Multiple Dependent Tests," Genetics, 166, 611-619.

Grapes, L., Dekkers, J. C., Rothschild, M. F., and Fernando, R. L. (2004), "Comparing Linkage Disequilibrium-Based Methods for Fine Mapping Quantitative Trait Loci," Genetics, 166, 1561-1570.

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011), "Extension of the Bayesian Alphabet for Genomic Selection," BMC Bioinformatics, 12, 186.

Kim, J. J., and Georges, M. (2002), "Evaluation of a New Fine-Mapping Method Exploiting Linkage Disequilibrium: A Case Study Analysing a Qtl with Major Effect on Milk Composition on Bovine Chromosome 14," Asian-Australasian Journal of Animal Sciences, 15, 1250-1256.

MacLeod, I. M., et al. (2010), "Power of a Genome Scan to Detect and Locate Quantitative Trait Loci in Cattle Using Dense Single Nucleotide Polymorphisms," J Anim Breed Genet, 127, 133-142.

Meuwissen, T. H., and Goddard, M. E. (2001), "Prediction of Identity by Descent Probabilities from Marker-Haplotypes," Genet Sel Evol, 33, 605-634.

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001), "Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps," Genetics, 157, 1819-1829.

Perez-Enciso, M., and Misztal, I. (2011), "Qxpak.5: Old Mixed Model Solutions for New Genomics Problems," BMC Bioinformatics, 12, 202.

Sun, X., Habier, D., Fernando, R. L., Garrick, D. J., and Dekkers, J. C. (2011), "Genomic Breeding Value Prediction and Qtl Mapping of Qtlmas2010 Data Using Bayesian Methods," BMC Proc, 5 Suppl 3, S13.

Uleberg, E., and Meuwissen, T. H. (2010), "Fine Mapping and Detection of the Causative Mutation Underlying Quantitative Trait Loci," J Anim Breed Genet, 127, 404-410.

Xu, S. (2003), "Estimating Polygenic Effects Using Markers of the Entire Genome," Genetics, 163, 789-801.

Zhao, H. H., Fernando, R. L., and Dekkers, J. C. (2007), "Power and Precision of Alternate Methods for Linkage Disequilibrium Mapping of Quantitative Trait Loci," Genetics, 175, 1975-1986.