Application of a three-haplotype LDLA model to the French Holstein population

David Jonas, Chris Hoze, Didier Boichard, Pascal Croiseau

Application of a Three-haplotype LDLA Model to the French Holstein Population

D. Jónás[1,2,3], C. Hozé[1,3], D. Boichard[1] and P. Croiseau[1]
[1]INRA, UMR1313 GABI, Jouy-en-Josas, France, [2]AgroParisTech, Paris, France, [3]UNCEIA, Paris, France

ABSTRACT: A new linkage analysis-linkage disequilibrium model was constructed with the aims of correction for long-range linkage disequilibrium and improve separation of closely linked QTLs. H1 hypothesis tested 3 QTLs vs. 2 for H0. Its performance, demonstrated earlier on a simulated dataset, was investigated here in a real-life situation. A French Holstein-Friesian bull population (n = 3940) with phenotypic records on milk protein content and genotypic records from 50,000 SNPs was used for the analysis. With this LDLA model, QTL localization improved significantly and we were able to distinguish 2 closely linked QTLs located on chromosome 20, namely the *GHR* gene and the *PRLR* gene, located only 6.9 Mb upstream from the *GHR*. Similar results were obtained after the analysis of 3 other chromosomes from the bovine genome. In addition, the width of the peaks also decreased considerably, resulting in narrower QTL predictions.
Keywords: LDLA; QTL mapping; haplotype analysis

Introduction

Due to the presence of complex pedigrees and large families in livestock populations, neither linkage analysis nor genome-wide association studies, by themselves are appropriate for QTL detection in animal populations with agronomical importance. A combined linkage analysis–linkage disequilibrium mapping method (LDLA) can (unlike simple linkage analysis methods) take full advantage of high-density marker genotypes and integrate historical recombination information into the analysis, as well as effectively incorporate complete population structure information (unlike genome-wide association studies; for an example, see Farnir, 2000). However, even with LDLA analysis methods, it is difficult to accurately differentiate and distinguish closely linked QTLs and in most of the cases a "ghost" QTL is detected between the true QTLs due to the presence of genetic markers that are linked to both real QTLs. In addition, the presence of long-range linkage disequilibrium (LD) also adds unwanted noise to the localization of the identified QTLs and leads to uncertain QTL positioning.

With the introduction of the composite interval mapping theory (Zeng, 1993) to LDLA mapping, we aimed to deal with the above mentioned limitations of LDLA mapping methods (Jonas et al., 2014). Namely, our main aim was to fine-map QTLs that were previously detected, but their true position could not be precisely predicted due to the long-range LD present in the examined populations. Secondly, we also aimed to improve the localization and differentiation of closely linked QTLs in livestock populations. Herein we present a real data application of the combined composite interval mapping-LDLA method to investigate the performance of the model in a real-life setting as well as to demonstrate the advantages of the new LDLA model and its gains in terms of QTL localization.

Materials and Methods

Data. Single nucleotide polymorphism (SNP) information was used to analyze four of the *Bos Taurus* autosomal chromosomes (BTA3, BTA5, BTA15 and BTA20), out of which the results of BTA20 are presented in details. The UMD-3 genome assembly of the bovine genome (Zimin et al., 2009) was used during the analysis for the localization of genomic markers and genes. Genotyping of individuals was done with the Illumina 50,000 SNP array (50K SNP-chip), while phasing of the SNPs was performed by the DagPhase software (Druet and Georges, 2010). In total, 1359 SNPs were mapped to BTA20, out of which 300 SNPs were removed prior to the analysis either due to low minor allele frequencies (applied MAF threshold: 5%) or because they were not in Hardy-Weinberg equilibrium ($p < 10^{-4}$). After the exclusion of markers from the extremities of the chromosomes due to the incompatibility of the model at these sites (explained in details below), 1045 SNPs were retained on BTA20 for the final analysis. For BTA3, BTA5 and BTA15 1812, 1503 and 1195 SNPs were retained, respectively. The average distance between the markers varied between 0.066 and 0.078 Mb for the different chromosomes. Haplotypes of 6 SNPs were constructed for the analysis.

Phenotype, pedigree and genotype records were available for the French Holstein Friesian population. Phenotypic performances were Daughter Yield Deviations of 3940 individuals for milk protein%. All of these animals were also genotyped with the Illumina 50K SNP-chip. In total, 12,142 animals were present in the available pedigree files, which were used to recover family relationships between the animals.

Model. A modified version of the LDLA model published by Meuwissen et al. (2002) is reported by Jonas et al. (2014) and it is briefly summarized below. Haplotype-cofactors are added to the original model in order to mask the long-range LD during the analysis. The milk protein% of the cows was modeled with the following equation:

$$y = \mu 1 + Z_u u + Z_{h_1} h_1 + Z_{h_2} h_2 + Z_{h_3} h_3 + e,$$

where **y** is the vector of observations, μ is an overall mean effect, **u** is a vector of polygenic effects, $\mathbf{h_1}$-$\mathbf{h_3}$ are vectors of random haplotype effects and **e** is a vector of random residuals. Z_u and $Z_{h_1}-Z_{h_3}$ are incidence matrices relating the random polygenic and the h_1-h_3 haplotype effects to the observations, respectively. This equation was used as the H1 alternative hypothesis in a likelihood ratio test (LRT), where the null H0 hypothesis model was

similar, but omitted the h_2 haplotype effect (which acted as the tested haplotype in the test):

$$y = \mu 1 + Z_u u + Z_{h_1} h_1 + Z_{h_3} h_3 + e,$$

where all symbols are defined identically as for the H1 hypothesis model. Under the null hypothesis, the distribution of the LRT values followed a chi-square distribution with 2 degrees of freedom. An equal distance of 3 cM was used between the three haplotypes for all chromosomes. The ends of the chromosomes within 3 cM were omitted from the current analysis, because the construction of the new LDLA model (abbreviated below as lrLD model) was not possible in these regions. Tests conducted on simulated data showed significant improvement of the results, as compared to those obtained by the application of the original 1-QTL model (Jonas et al., 2014).

An analysis based on the original model by Meuwissen et al. (2002) was also conducted and is referred to as MG model below. The results from the new LDLA model were evaluated in comparison to the MG model.

QTL detection. A Bonferroni-correction was implemented during the analysis (α = 1%) in order to determine significant QTL positions. A QTL was called at any location on the tested chromosomes, if the p-value of the specific location exceeded the Bonferroni-corrected α and this p-value was the highest within a +/-3 cM window around the position. A score was calculated as the ratio of the –log(p-value) at the peak's highest point vs. the average –log(p-value) in a +/-3 Mb window around the peak in order to describe the shape of the peak. The higher scores represented narrower peaks.

Results and Discussion

In order to test the lrLD model's performance on a real dataset, we selected the BTA20 chromosome, where a major QTL underlying milk protein content (growth hormone receptor; *GHR*) was identified previously in a Holstein-Friesian population (Blott et al., 2002). This chromosome was also a good example to test if the model is able to distinguish closely linked QTLs, since a second QTL (the prolactin receptor gene; *PRLR*) was located just 6.9 Mb upstream from the *GHR* gene. Quantitative trait locus' linked to milk protein% could be also hypothesized on the other 3 selected chromosomes based on an analysis conducted in our group (unpublished data).

Figure 1 shows the strength of association between the haplotypes of BTA20 and protein% for both the MG (grey) and lrLD (black) models. Predicted QTLs from the lrLD model (as described in the Materials and Methods section) are represented by red dots on the plot. The solid horizontal red line indicates the log-transformed, Bonferroni-corrected p-value threshold (α = 1%), while the two vertical dashed lines on the first plot indicate the region that is enlarged on the second plot of Figure 1. The 2 bars at the top of the 2nd plot indicate the positions of the *GHR* and the *PRLR* genes.

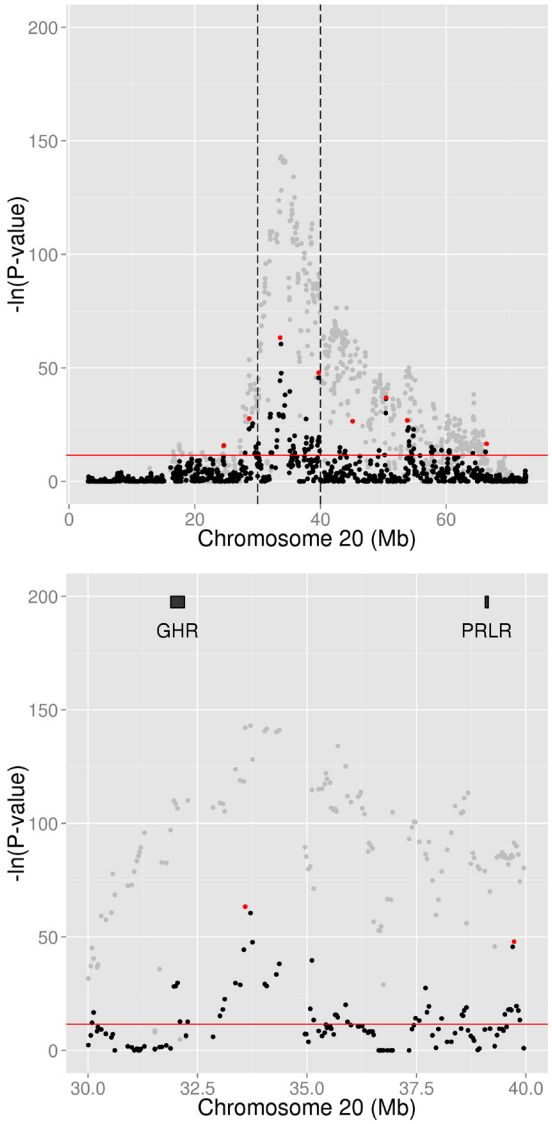

Figure 1: Genomic linkage map of BTA20. The transformed p-value threshold after Bonferroni-correction is represented by the solid red line (α= 1%). Grey dots show the results of the MG model, while black dots show those of the lrLD model. The red dots indicate the QTLs (as defined in the Materials and Methods section) detected with the lrLD model. The dashed lines on the first plot represent the *GRH-PRLR* gene-region and it is enlarged on the 2nd plot.

The p-values decreased considerably along the region of interest with the lrLD model: while 535 out of the 1045 tested positions were significant with the MG model, the number of significant associations between haplotypes and the examined trait decreased to 93 with the lrLD model. In addition, the width of the peaks decreased as well, which allowed the identification of 5 major and several smaller statistically significant peaks under the approximately 30 Mb broad peak, that was detected by the MG model. Out of the 5 putative QTLs, 2 were located within our region of interest: one between 33.01-34.37 Mb and the other one between 39.52-39.86 Mb. These peaks were 1.39 and 0.59 Mb far from the *GHR* and *PRLR* genes, respectively.

Although several genes are also located in the genomic regions covered by these peaks, none of them were related to milk protein content before. One of the smaller peaks at 31.96-32.10 Mb was located exactly under the *GHR* gene.

In addition to BTA20, we also tested the lrLD model on three other chromosomes of the bovine genome. In general, less QTLs were predicted with the lrLD model on most of the chromosomes (Table 1). The only exception from this was the BTA20 chromosome, where 8 QTLs were predicted with the lrLD model in contrast to the 4 predictions with the MG model. This is due to the extensive LD region around the *GHR* gene, where the detection of further QTLs were not possible with the MG model (Figure 1), however after successfully masking the long-range LD with the lrLD model, several additional QTLs (n = 4) could be detected in the region. With the latter model, the detected QTL-regions narrowed down by a significant amount in all cases (see Figure 1 for an example), which is also indicated by the average peak scores calculated for each chromosome in Table 1. The difference in the scores is especially conspicuous in case of BTA20.

Table 1: Number of predicted QTLs per chromosome with the original model published by Meuwissen et al. (denoted as MG) and with the model incorporating the haplotype cofactors (denoted as lrLD). The average peak score per chromosome is also indicated in the table.

Chromosome ID	Number of predicted QTLs		Average peak score	
	MG	lrLD	MG	lrLD
BTA3	8	5	4.61	5.98
BTA5	9	2	3.46	6.37
BTA15	9	3	5.03	8.98
BTA20	4	8	2.66	5.81

Conclusion

With the introduction of the haplotype cofactors to the LDLA model, the QTL localization improved significantly. The main advantages of the new model are the more precise estimation of the QTL locations and the larger power to distinguish closely linked QTLs.

Based on the presented real-data applications of the model, lower optimal distances were observed between the haplotypes in the model with respect to QTL localization, as compared to the optimal distance observed with the simulated dataset (Jonas et al., 2014).

The main limitations of the lrLD model are the relatively long running time and high computational power demand. However, since the primary benefit of the model is in the fine-mapping of QTLs, the model's drawbacks have only a limited effect on the analysis if the regions of interest are chosen in advance based on a prior QTL-analysis and the lrLD model is run only for the selected chromosomes or chromosome segments.

Literature Cited

Blott, S., Kim, J. J., Moisio, S. et al. (2003). *Genetics* 163: 253–266.

Druet, T., Georges, M. (2010). *Genetics* 184: 789-798.

Farnir, F. (2000). *Genome Res.* 10:220–227.

Jonas, D., Hozé, C., Boichard, D. et al. (2014). *Genet. Sel. Evol.* (submitted)

Meuwissen, T. H. E., Karlsen, A., Lien, S., et al. (2002). *Genetics* 161:373–379.

Zeng, Z. B. (1993). *Genetics* 136:1457–1468.

Zimin, A. V., Delcher, A. L., Florea, L. et al. (2009). *Genome Biol.* 10: R42.