QTL Detection for Fat Yield on BTA14 Using Linkage Disequilibrium Based Methods

M. J. Beribe, H. A. Carignano, N. Lopez-Villalobos, M. A. Poli and D. L. Roldán

1Instituto de Genética CICYCyA-INTA Castelar, Buenos Aires, Argentina, 2Massey University, Palmerston North, New Zealand.

ABSTRACT: In this study five models were used to identify quantitative trait loci (QTL) for fat yield in bovine chromosome 14 (BTA14): (1) a linear mixed model that includes covariance among all individuals, (2) the same linear mixed model with a probability that one individual belongs to one previously detected subpopulation as covariable, (3) the same mixed model association using the approach of principal component analysis, (4) a regression model (LD decay), and (5) a regression linkage disequilibrium linkage analysis (LDLA). The mapping population consisted of 23 paternal half-sib families with 870 offspring (77% Holstein purebreeds and 23% Jersey x Holstein crossbreds). All models detected a chromosomal region falling within the region that encompasses the postulated DGAT1 gene. In addition, both LD decay and LDLA identify two additional regions (25 Mbp and 80 Mbp) that could be associated to fat yield.

Keywords: dairy cattle, linkage disequilibrium, QTL

Introduction

Molecular genetic markers associated with phenotypic variation of a trait assist with the identification of genes underlying this trait. However, only a few causative genes associated with the variation in milk production have so far been reported (Bouwman et al. (2012)). Bovine chromosome 14 (BTA14) has been one of the most widely studied chromosome for quantitative trait loci (QTL) related to economically important traits in cattle (Wibowo et al. 2008; Ogorevc et al. 2009; Wang et al. 2012)). In dairy cattle, since the availability of SNP arrays, the genome-wide association analysis has provided a new tool in the search for the biologically causative genes. Nevertheless, the population stratification is an important problem mainly in association analysis in a livestock population. Several methods have been developed to deal with unknown population structure. Among these methods, structured association (Pritchard et al. 2000), and principal component analysis (Price et al. 2006)) are the most widely used. Structured association provides direct information about population structure with an easy interpretation, but it can lose power when too many subpopulations are specified. Principal component analysis is a linear dimensionality reduction technique used to infer continuous axes of genetic variation. For the animal population where the family relationship is strong, combined linkage disequilibrium and linkage analysis (LDLA) has been suggested as the method to enhance the feasibility of fine-mapping QTL (Meuwissen et al. 2002)).

This paper reports the results of association analysis between SNPs located in the BTA14 chromosome and fat yield using two different analysis methods, association and LDLA approaches.

Materials and Methods

Animals and phenotypic data. A total of 339,730 herd-test records for daily yields of milk and fat and protein percentages were collected from 1998 to 2012 from 1999 cows distributed in 37 commercial herds located in the central-east area of Argentina (76% purebred Holstein and 24% Holstein x Jersey crossbred cows). Daily yields of fat and protein were derived from milk yield and fat and protein percentages. Lactation yield of fat adjusted at 305 days in milk for each cow was estimated using a random regression model with SAS (2010). The random regression model included the fixed effect of proportion Holstein as a covariable, lactation number (9 levels), herd-test date (32 levels) and calving season (4 levels). The random effect of cow during the lactation (days in milk) was modeled with a fourth degree Legendre orthogonal polynomial. The (co)variance components of the random regression coefficients were estimated assuming an unstructured matrix, and residual errors were assumed to be independent with a common variance across the lactation. Only fat yield of the first lactation from 925 cows was retained for the QTL detection.

Molecular data. All animals were genotyped with the Illumina BovineSNP50 BeadChip with 54,609 SNPs. To ensure the quality of the genotypes, control parameters were chosen, i.e. minor allele frequency (MAF) and SNP and individual call rates. In addition parents/offspring inheritance discordancess were checked based on inheritance Mendelian rules.

The QTL study was carried out with 670 purebred Holstein cows and 200 Holstein x Jersey crossbred cows, which were the progeny of 21 Holstein sires and two Jersey sires. The average number of offspring per sire was 35.7; a minimum of 2 (3 cases) to a maximum of 119 (1 case) family size.

QTL models. Five models were evaluated, including mixed and fixed models. The first model (denoted as AA) was a linear mixed model approach that includes correlation among all individuals. The statistical model can be expressed as: \( y = X\beta + Zu + e \), where \( y \) is a vector of phenotypes for FY, \( \beta \) is the vector of SNP effect, \( X \) is a design matrix of SNP genotypes, \( u \) is the random additive effect, \( Z \) is a incidence matrix for a random additive effects
demonstrated a chromosome-wise p-value <1.76x10^{-3}, or a
We considered a QTL to be genome-wise significant if it
were repeated by including the genotype of the SNP as a
principal components analysis were included in the first
model as covariables. These three association analyses were
performed using the GCTA computer programme that uses
REML procedures for solving the mixed model equations
(Yang et al. (2011)).

The fourth (denoted as LD) and fifth (denoted as
LDLA) models were based on association and joint linkage
and association analyses (Legarra and Fernando (2009))
The LD model perform association analysis based on a LD-
decay method where parental haplotypes are pooled in
classes, defined by the haplotype IBS status. To a given
class corresponds a specific effect on the trait. This model
is expressed as: \( y = \mathbf{m} \beta + Z_p \mathbf{W}_p \mathbf{Q} \mathbf{m} \beta + Z_m \mathbf{Q} m \beta + e \), where \( y \)
are the phenotypes which are conditionals on all marker
information \( m \) and on the haplotype substitution effects \( \mathbf{p} \). \( W_p \)
are the transmission probability for a QTL allele from sires
to paternal chromosome in the offspring, \( Q_S \) and \( Q_m \) are incidence matrices relating haplotypes in the sires,
and maternal haplotypes in the offspring, to appropriate
elements in \( \beta \). \( Z_p \) and \( Z_m \) are appropriate incidence matrices relating paternal and maternal gametes in the
progeny to records. The LDLA model also performs an
association analysis based on a joint linkage and linkage
disequilibrium method under the following model: \( y = \mathbf{m} \beta + Z_p \mathbf{W}_p \mathbf{Q} \mathbf{m} \beta + Z_m \mathbf{Q} m \beta + Z_p \mathbf{W}_p v + e \), where \( v \)
are the within
sire QTL effects. The QTLMAP software (Elsen et al.
(1999)) was used to perform both LD and LDLA analysis.
The haplotype length was set to 4 SNP and the genome
scans were performed using a 0.1 cM step (about
1Mbp-1cM). Chromosome-wise significance thresholds
were determined by permutation from the experiment by
1,000 iterations (Churchill and Doerge (1994)).

Three genome-wise level thresholds were used.
We considered a QTL to be genome-wise significant if it
demonstrated a chromosome-wise p-value <1.76x10^{-3}, or a
p-value <3.46x10^{-4}, or a p-value <3.45x10^{-5} as these values
correspond to a genome-wise p-value of 5%, 1% or 0.1%
respectively (chromosome-wise p-value × 29 autosomes).

For the SNP showing genome-wise significant
association encompassing the DGAT1 gene, the analyses
were repeated by including the genotype of the SNP as a
fixed effect in the QTL model. These analyses were
performed with the aim to assess the possible direct
relationship of the tested SNP with the QTL effect
identified.

Results and Discussion

Marker information. Monomorphic or ungenotyped SNPs were removed. A subset with unknown
location in the Bta UMD3.1 assembly was also discarded.
In addition, SNPs with a call rate < 90 %, with a MAF<1%,
or with a significant (p<1.0x10^{-8}) deviation from the Hardy-
Weinberg equilibrium were omitted for subsequent
analysis. Finally, 44,163 informative SNP (43,327 in the 29
autosomes) out of the 54,609 originally genotyped SNPs
remained for analysis. The average SNPs by chromosome
was 1,500 SNPs, with a minimum number of markers on the
BTA27 (n=807) and a maximum on the BTA1 carrying
2,804 SNPs. SNP density was in a 18 to 23 SNPs/cM range,
whereas the heterozygosity ranged around 29.50% (BTA15
and BTA20) to 31.80% (BTA23).

Only individuals with a call rate >90% were retained. Nevertheless, one Jersey sire with a call rate
slightly lower than this threshold (84.40%) was retained.
Globally, the sires had a call rate beyond 98.30%, and every
genotyped offspring had a call rate around 99.2%. A total of
55 animals showing major inconsistencies between
pedigree and genomic relationship were excluded from the
analyses.

QTL for dairy traits. Here we present QTL
results from fat yield in the first lactation on BTA14.
Trying to provide a global overview of the comparison of
the results obtained through the methodologies described
above, a summary table is provided in Table 1.

All AA mixed models detects the same significant
SNP (rs109421300 at 1.801116 Mbp). Wang et al. (2012)
reported the same marker as the most significantly
associated SNP (p=1.54x10^{-198}) to milk fat percentage in a
German Holstein population. They performed a genome-
wide association study accounting the population structure
by adjusting the relationships between individuals by a
genomic relationship matrix.

In case of family structures or a mixture of breeds,
some methods correcting for population stratification as
structured association methods or principal component
analysis based approaches should be applied. Our results
from a real data suggest that considering either structure
population or principal component covariables perform well
in the AA mixed model case, and the later AA pca seems
slightly better that AAstr (according to p-values, data not
shown).

Globally, three chromosome regions associated
with FY were identified for LD and LDLA methods on at
least one threshold significance level. These regions are
localized from 4 to 12 Mpb, 24.5-25.5 Mpb and ~ 80 Mbp
(Table 1). The region around 0–10 Mbp has many QTL
reported associated to the milk production traits. In
particular, mainly clustered in a region of 0.26 to 3.81 Mpb
(Grisart et al. (2002); Winter et al. (2002)), the DGAT1
gene is well known to have a large effect on milk FP and
FY in dairy cattle (Grisart et al. (2002); Winter et al.
(2002); Thaller et al. (2003); Bennewitz et al. (2004)). In
the same region, the CYP11B1 gene (from 2.770329 to
2.775442 Mbp) was reported negatively associated with
MY and FY (Kaupe et al. (2007)). Additional QTL for FY
were identified at 24.67Mbp-27.34 Mbp (Harder et al.
(2006); Ogorevc et al. (2009)) and at 51.26 and 79.76Mbp.
However no candidate genes have been explored for these
Table 1. QTL detected in at least one association and joint linkage and linkage disequilibrium analyses (AA, AAstr, AApca, LD, and LDLA)§

<table>
<thead>
<tr>
<th>Position (Mbp)</th>
<th>Model</th>
<th>AA</th>
<th>AAstr</th>
<th>AApca</th>
<th>LD</th>
<th>LDLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.69</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.80</td>
<td></td>
<td>***</td>
<td>***</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.05</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.38</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.52</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.46</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.36</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.16</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.55</td>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.26</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>(*)</td>
</tr>
<tr>
<td>24.06</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.43</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.46</td>
<td></td>
<td>**</td>
<td>(*)</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.56</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.66</td>
<td></td>
<td>***</td>
<td>(*)</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>***</td>
<td>(*)</td>
</tr>
<tr>
<td>24.86</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>***</td>
<td>(*)</td>
</tr>
<tr>
<td>24.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>***</td>
<td>(*)</td>
</tr>
<tr>
<td>25.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>(*)</td>
</tr>
<tr>
<td>25.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51.26</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>79.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>**</td>
<td>(*)</td>
</tr>
</tbody>
</table>

§ AA=Association mixed model with a genomic based relationship matrix; AAstr=Association mixed model with a genomic based relationship matrix with the probability that one individual belongs to one subpopulation as covariable; AApca: Association mixed model with a genomic based relationship matrix with the four first eigenvectors from principal components analysis as covariable; LD= Association analysis based on LD decay model. LDLA= Joint linkage and linkage disequilibrium analysis.

Significance threshold levels. *P<0.05 genome-wise significant threshold; **= P<0.01 genome-wise significant level; ***= P<0.001 genome-wise level. In parentheses: significant threshold level when the most significant SNP was adjusted as fixed effect in the QTL model.

Only the first region from LD or LDLA methods has exhibited consistent findings with the only significant SNP detected with the AA mixed models.

Further analysis performed by fitting as a fixed effect in the corresponding QTL model the significant SNP identified in the mixed models (SNP rs109421300). By repeating the LD and LDLA analyses with this SNP fixed, the significance of FY QTL decreased from the 0.1% genome-wise significance level to 5% genome level (Table 1). Therefore the remained significant association with FY was identified only on the region from 24.5 Mbp to 25.5 Mbp obtained using the model without the fixed SNP (Table 1). The purpose of fitting a regression was to exclude spurious effects due to linkage with another QTL of large effect. The fact that for these chromosomal regions affecting FY the fixation of the SNP selected based on the association analysis resulted in a substantial decrease of the QTL significance suggests that the fixed SNP could be in linkage disequilibrium with the causal mutation but not directly responsible of the QTL effects.

Conclusion

The comparison of the results reported in the current study, demonstrate that the LDLA method identify several map positions not detected in the linkage disequilibrium based models. As has been suggested previously, these results could indicate that the LDLA method may provide an important improvement in the power and efficiency of QTL detection in half-sib experimental designs. Nevertheless, the comparison of the association and the joint linkage and association analysis methods seems a good approach for obtaining a global picture of all identifiable QTL. Further work would be necessary to decipher the underlying genetics mechanisms.

Literature Cited