An Information Measure For Founder Haplotype Informativeness Useful For QTL Fine Mapping Experiments

C. Baes*[+], *M. Mayer*[+], J. Bennewitz* and N. Reinsch[+]

Introduction

Fine mapping quantitative trait loci (QTL) in segregating populations genotyped for a dense set of markers often relies on a combination of linkage analysis (LA) and a linkage disequilibrium (LD) mapping (LDLA mapping, Meuwissen et al. 2002). LA uses meioses that occurred within the genotyped generations of individuals. LA mapping is based on long range linkage disequilibrium within families, which makes it robust, but limits the precision. LD mapping uses historical meioses (i.e. all meioses since the causal mutation was assumed to be occurred). Here population wide disequilibrium is utilised, which is usually limited to small chromosomal regions.

When using LDLA mapping marker haplotypes are divided into those that are inherited from genotyped ancestors (descendent haplotypes) and those that are inherited from individuals outside the genotyped pedigree (founder haplotypes). In application of LDLA mapping in dairy cattle populations, often the sires are genotyped, but not the dams. In these situations, the founder haplotypes are the haplotypes of the founder sires and those haplotypes of genotyped individuals, which they inherited from their dam. Hence, for a given sire, the origin of both haplotypes (i.e. paternally or maternally derived) and the constitution of alleles on the haplotypes have to be known or estimated. The quality of estimation is, however, not equal for all chromosomal positions and depends on the predefined length of haplotype, marker density, and marker informativity.

In this article we describe an entropy-based information measure for the founder haplotype informativeness, especially for the maternal haplotypes, indicating how precise the allelic constitution of maternal haplotypes can be reconstructed from data. Subsequently we show that the test statistic of an LDLA approach can be affected by different information contents. Finally, some recommendations are given.

Notation and Definitions

Genotype and Haplotype. A series of genetic markers are genotyped at L loci for N individuals. Genotype is defined as the information at L marker loci. For each individual $i=1,\ldots,N$, $g_i=(g_{i1},\ldots,g_{iL})$ represents the individual's genotypes at a given locus l, $l=1,\ldots,L$. An example genotype for an individual i at 3 loci is $g_i=(Aa,Bb,Cc)$. A haplotype is defined as the phased multi-locus information across L loci of a given chromosome. An example haplotype for individual i is $h_i=(ABC)$. The haplotypes is either paternally or maternally derived.

* *Universität Hohenheim, Institut für Tierhaltung und Tierzüchtung, D-70593 Stuttgart, Germany*
+ *Leibniz-Institut für Nutztierbiologie (FBN), FB Genetik und Biometrie, D-18196 Dummerstorf, Germany*

Diplotype. The diplotype is a set of haplotype pairs which correspond to a given genotype. There may be several possible haplotype pairs for a given genotype. The ordered diplotype is a set of ordered (ordered by parental origin) haplotype pairs, where the first haplotype (h_{i1}) is defined as paternal and the second (h_{i2}) as maternal. For an individual *i* with genotype g_i*=(Aa,Bb,Cc)* the diplotype is d_i*=[(ABC,abc), (AbC,aBc), (Abc,aBC), (ABc,abC)]*. The ordered diplotype is d_i*=[(ABC,abc), (AbC,aBc), (Abc,aBC), (ABc,abC), (aBC,Abc), (aBc,AbC), (abC,ABc), (abc,ABC)]*.

Haplotype windows. It is assumed that a putative QTL is located within a marker interval. There are *Q* (*Q*=*L*-1) different possible QTL positions. Each *Q* is surrounded by a 'sliding window' with at least one marker left and one marker right of the QTL. Window size is given in cM which allows for flexibility if the markers are not distributed evenly. The window is therefore a short subsection of the full haplotype.

Entropy and information content measure. In physics and information theory, Shannon entropy S (Shannon 1948) is a proper measure to quantify the non-order or the degree of non-structure of a system. It is defined as $S = \sum - p_j \log_2 p_j$, where p_j denotes the probabilities of the states the system can enter and the summation is over all possible states. S therefore reaches its maximum when the system exhibits the lowest degree of structure, which occurs when all states are equally probable.

A set of ordered diplotypes of length *L* that contains all possible ordered diplotypes of an individual can also be considered a system with the possible ordered haplotypes as its states. The sum of all possible haplotype probabilities (the probability vector for a given haplotype) is equal to one. In the absence of genotype information from relatives, the haplotype probabilities of all possible haplotypes of an individual are equal and the system exhibits the lowest degree of structure (S is maximal). In this case the haplotype probabilities at a given chromosomal position using a defined window size are $p(hap_j)_{pos} = 1/k_{pos}$, where *k* is the number of possible haplotypes within the window at position *pos.* This state is considered 'naïve', and the naïve entropy (E_o) is calculated as $E_0 = \sum_{j=1}^{k} - p(hap_j)_{pos}) \times \log_2 p(hap_j)_{pos}$. Entropy is maximal in this state, as all haplotypes are equally likely. If genotypic information of relatives is used, the true entropy is calculated as $E_T = \sum_{j=1}^{k} - p(hap_j \mid g)_{pos}) \times \log_2 p(hap_j \mid g)_{pos}$. Here $p(hap_j \mid g)_{pos}$ denotes for the haplotype probability of haplotype *j* given genotype information of relatives (*g*) at position *pos*. The information content for this individual is expressed as $I_i = 1 - E_{T_i} / E_{0_i}$.

Materials and Methods

We applied the concept of entropy information content measure to previously published data. Briefly, LDLA mapping was performed in dairy cattle with genotyped male half sib families. Genotypes were collected on BTA02, BTA18 and BTA27, the marker density was more or less dense. For a more detailed description see Baes et al. (2009, 2010a,b). The paternal origin of haplotypes was determined using informative markers only. We argue that the information on this path is high if dense markers are used. Furthermore, because of the long range linkage disequilibrium within families, the mapping results (produced by LA within the combined LDLA approach) are not affected by varied information content on this path. Therefore we concentrate on the information content for the determination of the maternal derived haplotypes. Here we used the sliding window approach described above. For a certain window the set of possible maternal haplotypes of a sire are derived using the full length haplotype as starting information. Here *k* is the number of possible maternal haplotypes. The probability of each haplotype within the window is estimated using marker information of the sire and its relatives. Following this, the information content averaged over all individuals at a chromosomal position using a defined window size is expressed as

$$I_{ave} = \frac{1}{N}\sum_{i=1}^{N} I_i \text{ , with } I_i = 1 - \frac{E_{T_i}}{E_{0_i}} .$$

Results and Discussion

For detailed mapping results see Baes et al. (2009, 2010a,b). Here we concentrate on the maternal haplotype information content and its impact on likelihood test statistic profiles. For BTA02, test statistics were lower than those calculated using simple LA (Baes et al., 2010a). This may be explained by the lack of LD found in the experiment. Notably, the likelihood profile of the LA was much smoother than the profile calculated with combined LDLA analysis using variance components, despite the lack of LD information. This phenomenon may be explained when the information content of the maternal haplotypes is examined (Figure 1, BTA02). The likelihood profile follows the profile of the information content quite closely, meaning the likelihood test statistics calculated using the haplotype approach are quite sensitive to the amount of information on maternal haplotypes. Although the information content is generally quite high (average = 87.9%), the information content of maternal haplotypes within the proximal 50 cM of the chromosomal region under investigation is higher than that of the distal 30 cM. A similar picture can be observed at the beginning of BTA18 (Figure 1). Furthermore it should be considered that the LD-likelihood profile showed some evidence for LD in the region around 72 cM. In BTA27 the information content is constantly high (not shown), hence, this phenomenon was not observable (Baes et al. 2010b).

Conclusion

In this article we describe an information content measure for haplotype windows based on Shannon entropy. It can be computed for subsequent positions on a chromosome taking all markers within windows of defined length into account. When applied to maternal

haplotypes our information measure indicates how much LD information really can be utilized. It can be concluded that the information content may have an effect on the likelihood ratio values as discussed for our examples and should therefore be taken into account when interpreting fine mapping results obtained from LDLA mapping.

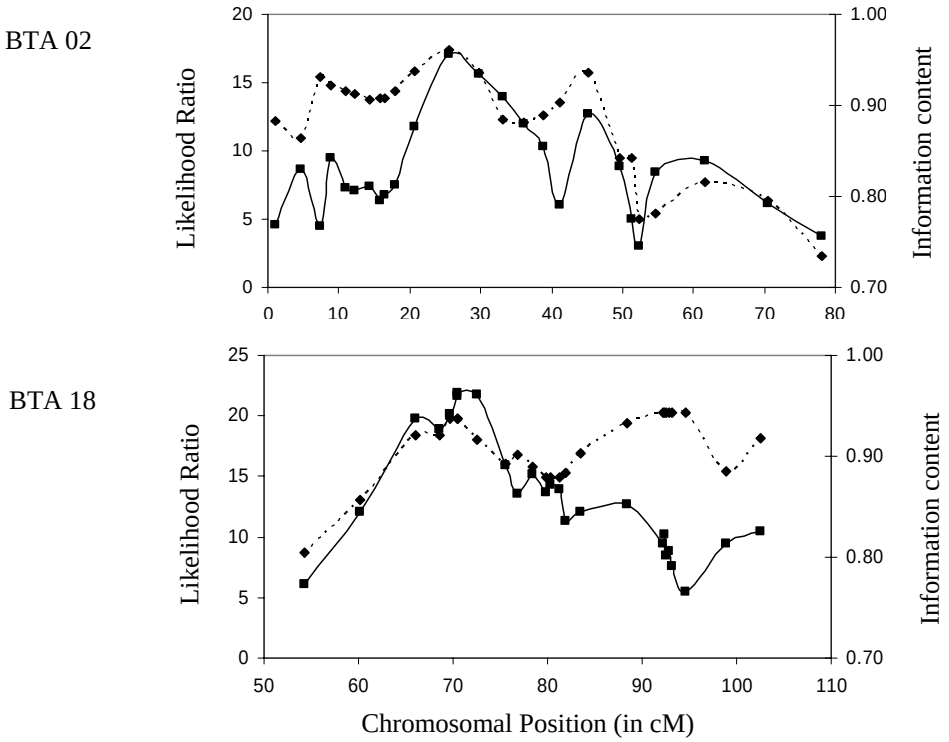

Figure 1: Test statistics obtained from LDLA QTL mapping (closed lines) and average maternal haplotype information content (dashed lines) for each chromosomal position on BTA02 (top) and BTA18 (bottom). The symbols indicate putative QTL positions.

References

Baes, C., Brand, B., Mayer, M. *et al.* (2009). *J. Dairy Sci.*, 92:4046–4054.

Baes, C., Görtz, I., Mayer, M. *et al.* (2010a). *J. Anim. Breed. Genet.*, in press (online early view).

Baes, C., Mayer, M., Tetens, J. *et al.* (2010b). *Can. J. Anim. Sci.*, in press.

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

Shannon, C. (1948). *Bell System Technical Journal*, 27: 379–423, 623–656.