Jacquin *et al. Genetics Selection Evolution* 2014, **46**:45
http://www.gsejournal.org/content/46/1/45

GSE Genetics Selection Evolution

RESEARCH Open Access

Using haplotypes for the prediction of allelic identity to fine-map QTL: characterization and properties

Laval Jacquin[1,2,3*], Jean-Michel Elsen[1,2,3] and Hélène Gilbert[1,2,3]

Abstract

Background: Numerous methods have been developed over the last decade to predict allelic identity at unobserved loci between pairs of chromosome segments along the genome. These loci are often unobserved positions tested for the presence of quantitative trait loci (QTL). The main objective of this study was to understand from a theoretical standpoint the relation between linkage disequilibrium (LD) and allelic identity prediction when using haplotypes for fine mapping of QTL. In addition, six allelic identity predictors (AIP) were also compared in this study to determine which one performed best in theory and application.

Results: A criterion based on a simple measure of matrix distance was used to study the relation between LD and allelic identity prediction when using haplotypes. The consistency of this criterion with the accuracy of QTL localization, another criterion commonly used to compare AIP, was evaluated on a set of real chromosomes. For this set of chromosomes, the criterion was consistent with the mapping accuracy of a simulated QTL with either low or high effect. As measured by the matrix distance, the best AIP for QTL mapping were those that best captured LD between a tested position and a QTL. Moreover the matrix distance between a tested position and a QTL was shown to decrease for some AIP when LD increased. However, the matrix distance for AIP with continuous predictions in the [0,1] interval was algebraically proven to decrease less rapidly up to a lower bound with increasing LD in the simplest situations, than the discrete predictor based on identity by state between haplotypes (IBS_{hap}), for which there was no lower bound. The expected LD between haplotypes at a tested position and alleles at a QTL is a quantity that increases naturally when the tested position gets closer to the QTL. This behavior was demonstrated with pig and unrelated human chromosomes.

Conclusions: When the density of markers is high, and therefore LD between adjacent loci can be assumed to be high, the discrete predictor IBS_{hap} is recommended since it predicts allele identity correctly when taking LD into account.

Background

Numerous methods have been developed to predict allelic identity at an unobserved locus between pairs of chromosome segments. Such predictions are generally carried out by observing allelic similarities between the pairs of chromosome segments that surround this locus [1-3]. It is assumed that chromosome segments that exhibit more similarities have a higher chance of harboring the same allele(s) at this locus. Many of these methods [1-5] use either directly or implicitly the concept of identity-by-descent (IBD), and therefore predict allelic identity based on allelic likeness. Such predictions of allelic identity can be either continuous or discrete in the [0,1] interval. The matrices that contain these predictions for pairs of chromosome segments, at an unobserved locus, can be used in a statistical procedure to detect association between the locus and some phenotypes of interest. For example, these matrices can be interpreted as being proportional to the

*Correspondence: Julien.Jacquin@toulouse.inra.fr
[1]INRA, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), F-31326, Castanet-Tolosan, France
[2]Université de Toulouse, INP, ENSAT, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), F-31326, Castanet-Tolosan, France
Full list of author information is available at the end of the article

covariance matrices of the effect of the locus on phenotypes of interest [1,4,6] and therefore play a central role in the statistical analysis of the variability. The similarity between chromosome segments can be measured based on the haplotypes of markers carried by the segments. Indeed, it has been shown that haplotype-based methods have a higher potential to detect trait-marker associations than single-marker methods in some cases [7-16]. Different methods for predicting allelic identity, hereafter named Allelic Identity Predictors (AIP), have been proposed and in this study, we have compared some of these methods i.e.: (1) the probability measure described by Meuwissen and Goddard [1] is the conditional probability of being IBD at an unobserved locus for pairs of haplotypes, given the identical-by-state (IBS) status of alleles spanning that position; (2) the similarity score of Li and Jiang [2] calculates the sum of the number of shared alleles and the length of the longest shared substring that spans an unobserved locus for pairs of haplotypes; (3) the probability model of Browning [3] is based on Variable Length Markov Chains (VLMC) and performs chromosome clustering at a given marker, and in this model, chromosomes that belong to a given cluster are considered as potentially harboring the same unobserved allele(s) locally; and (4) the IBS status of all marker alleles between pairs of haplotypes and (5) the IBS status of single markers alleles, which are the simplest AIP.

In some association studies, such as those that use random effect models for example, the only input that differs from one AIP to another is the similarity (covariance) matrix built for the tested location. Thus, investigating the properties of similarity matrices is another strategy when comparing AIP, since this comparison is generally based on the accuracy of quantitative trait locus (QTL) localization (e.g. root mean square error). The main objective of the present study was to understand the relation between linkage disequilibrium (LD) and allelic identity prediction when using haplotypes, by identifying the properties of similarity matrices in the neighborhood of a QTL and at the QTL. This was performed using a simple distance measure between these matrices and the similarity matrix at the QTL based on the observed allelic identity (IBS). This distance measure was expressed analytically in terms of LD coefficients. There has been an increasing interest in taking advantage of LD for fine-mapping of complex disease genes [17-20] and QTL [21-24]. Nevertheless, to the best of our knowledge, no study has yet used analytical methods to compare AIP in relation to LD. Here, we define a new criterion based on the chosen matrix distance measure, which allows discrimination between the six AIP. We evaluated the consistency of this criterion with the mapping accuracy of the six AIP for a QTL simulated according to different LD patterns and populations.

The simulations were based on two population types, a set of human chromosomes and a set of porcine chromosomes, with different LD and density patterns. In each case, the QTL was a hidden SNP that simulated a biallelic QTL, as previously proposed [4,8,23,25]. Hence, the present study was framed around the common idea that there is a favorable allele at the QTL, which affects an observed trait. In this context, the aim of AIP is to predict, at the QTL, whether both chromosomal segments of any pair harbor the same unobserved favorable allele or not, which is the same as predicting the IBS or non-IBS state of the alleles. A new (6) unreferenced AIP, named trained predictor and abbreviated as TP, is also compared in this paper. This new predictor, based on a matrix distance concept similar to the one used to discriminate between the AIP, performs least squares prediction in a global fashion over chromosomes. The purpose of this predictor was to investigate the behavior of an AIP which performs global training over the chromosomes in relation to local patterns of LD.

Methods

Matrix distance comparison

Let $\mathcal{I} = \{i_1, \ldots, i_r\}$ be a set of positions that are tested for the presence of a QTL on $2n$ phased homologous chromosomes for n diploid individuals. Only one QTL is considered to be in the screened region. In a sliding window approach, each position tested is considered to be the unobserved center of the haplotypes carried by different chromosome segments. Figure 1 shows an example of tested positions for a sliding window of six markers and a QTL located between SNPs 39 and 40.

Let $s^{\mathcal{P}}_{i,c_1,c_2} \in [0,1]$ be the IBS or IBD prediction of allelic identity, depending on an AIP $\mathcal{P}$ at a tested position $i \in \mathcal{I}$, for a couple (c_1, c_2) of chromosome segments. Note that $s^{\mathcal{P}}_{i,c_1,c_2}$ is calculated according to the observed similarity between the haplotypes carried by c_1 and c_2. Hence, c_1 and c_2 can harbor different unobserved alleles at i even if these segments carry the same haplotype. We define $\mathbf{M}^{\mathcal{P},i} = \left(s^{\mathcal{P}}_{i,c_1,c_2}\right)_{1 \leq c_1,c_2 \leq 2n}$ as the similarity matrix built from the predictions of allelic identity at locus i for $\mathcal{P}$. Matrix $\mathbf{M}^{\mathcal{P},i}$ can be used in a statistical procedure to detect association between i and some phenotype of interest.

Let $u^{QTL}_{c_1,c_2} \in \{0,1\}$ be the true allelic identity observed at the QTL (IBS) for a couple (c_1, c_2) of chromosome segments. On the basis of known alleles at the QTL, the similarity $\mathbf{M}^{QTL} = \left(u^{QTL}_{c_1,c_2}\right)_{1 \leq c_1,c_2 \leq 2n}$ can be built with the real allelic identities. Note that $\mathbf{M}^{QTL}$ is simply a similarity matrix that describes the IBS or non-IBS state of alleles at the QTL.

Let d_1 be a normalized distance measure between $\mathbf{M}^{\mathcal{P},i}$ and $\mathbf{M}^{QTL}$ induced by the entrywise 1-norm, which is the

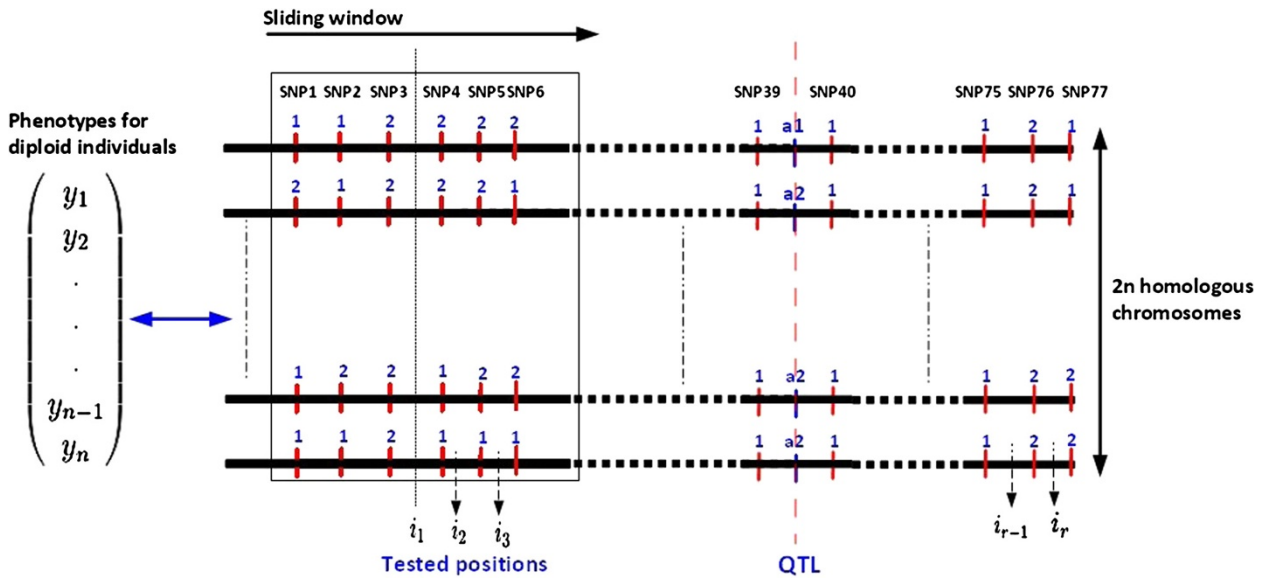

Figure 1 Principle of the sliding window approach when screening a chromosome for the set $\{i_1, \ldots, ir\}$ of tested positions for associations between phenotypes and 6-marker haplotypes.

sum of the absolute differences between the elements of two matrices or vectors, i.e.

$$d_1\left(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}\right) = \frac{1}{4n^2}\|\mathbf{M}^{\mathcal{P},i} - \mathbf{M}^{QTL}\|_1 = \frac{1}{4n^2}\sum_{c_1=1}^{2n}\sum_{c_2=1}^{2n}|s^{\mathcal{P}}_{i,c_1,c_2} - u^{QTL}_{c_1,c_2}|$$

Note that some AIP have continuous prediction errors $|s^{\mathcal{P}}_{i,c_1,c_2} - u^{QTL}_{c_1,c_2}|$ in $[0,1]$, while for others, prediction errors are limited to the discrete set $\{0,1\}$. Measure d_1 is therefore more appropriate than the euclidean metric d_2, for example, because it does not shrink continuous prediction errors in $[0,1]$. Let θ_{QTL} be the position of the QTL. When a predictor $\mathcal{P}$ performs well, $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ should be minimum at the tested position closest to θ_{QTL}. Hence $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ can be used to compare different AIP for a set of tested positions. Note that $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ can also be expressed as [see Additional file 1]:

$$d_1\left(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}\right) = \sum_{p=1}^{K} f_{i,h_p} \sum_{q=1}^{K} f_{i,h_q}\left[p^{QTL}_{i,h_p,h_q} \times \left(1 - s^{\mathcal{P}}_{i,h_p,h_q}\right) + \left(1 - p^{QTL}_{i,h_p,h_q}\right) s^{\mathcal{P}}_{i,h_p,h_q}\right] \quad (1)$$

where $K = 2^t$ is the number of possible observed haplotypes at position i, for a sliding window of t markers. f_{i,h_p} and f_{i,h_q} are the frequencies of haplotypes h_p and h_q at position i, respectively. Note that some haplotypes among the K possible haplotypes may not be observed in practice. Hence, the corresponding frequencies for these haplotypes will naturally be equal to 0 in expression (1). p^{QTL}_{i,h_p,h_q} is the proportion of identical alleles shared at the QTL by the pairs of chromosomes that carry h_p and h_q, at position i, and $s^{\mathcal{P}}_{i,h_p,h_q}$ is the prediction of allelic identity at locus i for the predictor $\mathcal{P}$ and a pair (h_p, h_q) of haplotypes. Expression (1) will be used subsequently to express d_1 as a function of LD coefficients, and to understand the trained predictor defined in this paper.

Measures of AIP evaluated

The AIP evaluated in this study were IBS_m (IBS status of alleles at single markers), IBS_{hap} (IBS status of all marker alleles between pairs of haplotypes), P(IBD) (IBD probability of Meuwissen and Goddard [1]), Score (similarity score of Li and Jiang [2]), Beagle (cluster-based probability model of Browning [3]) and TP (the trained predictor). Note that the tested positions coincide with marker positions for IBS_m and Beagle. These positions are therefore different from those in Figure 1. The tested positions for IBS_{hap}, P(IBD), Score and TP are defined as presented in Figure 1.

IBS_m gives an allelic identity prediction of 1 if a pair of chromosome segments carries the same allele at a tested marker and 0 otherwise. With IBS_{hap} the prediction of allelic identity is equal to 1 if both chromosome

segments of a pair carry the same marker alleles for haplotypes that span the tested position i, and 0 otherwise. P(IBD) is an estimation of the conditional probability of being IBD at i for a pair of chromosome segments, given the IBS status of marker alleles of the haplotypes spanning i. This measure of probability is based on a coalescence process and models recombination between markers. The P(IBD) function was applied here with an ancestral effective population size of 100 and 100 generations from the base population, as in Meuwissen and Goddard [1]. Meuwissen and Goddard [23] showed that violations of these assumptions, i.e. that alter the effective population size and the number of generations since the base population, had no effect on the mapping accuracy of their methods [23,26]. For a pair of haplotypes carried by two chromosome segments, Score is the summation of the number of IBS alleles and the length of the longest common substring of IBS alleles that span i. Score integrates weight functions that decrease the significance of markers based on their genetic distance from i. As proposed in Li and Jiang [2], these functions were chosen to be one minus the distance, in centiMorgan (cM), of each marker from i on the haplotypes within the sliding window (as presented in Figure 1). Beagle clusters chromosomes or haplotypes locally at a tested marker if they have similar probabilities of carrying the same alleles at following adjacent markers. The Beagle probability model was built at each marker by running the Beagle software (Beagle 3.3.2; http://faculty.washington.edu/browning/beagle/beagle.html, Browning [3], Browning and Browning [12]) and fitting all the chromosome markers at one time. The Beagle probability model needs two parameters (scale and shift) to be built. These parameters were first estimated from the data using a cross-validation procedure. However, the mapping results were less accurate than those obtained with the default values for these parameters that were proposed by the authors. According to the authors, the default values have performed well in simulation studies and real data analyses [12,27]. Hence the default values scale = 4.0 and shift = 0.2 [12] were used.

The trained predictor (TP), built by least squares prediction, is based on the idea that pairs of haplotypes that exhibit the same amount of allelic similarity should have the same probability of harboring identical alleles, regardless of the tested positions they span. Estimates for $\left(s^{TP}_{i,h_p,h_q}\right)_{(p,q)\in\{1,..,K\}^2}$ can be obtained as follows. Let $\mathcal{J} = \{j_1, \ldots, j_T\}$ be a set of observed SNPs on chromosomes, which are called target SNPs. Each target SNP j is defined as the middle marker of a sliding window of $t+1$ loci, where t is the number of observed flanking markers used to predict allelic identity at the target SNP. Let $u_{j,c_1,c_2} \in \{0,1\}$ be the real allele identity at j for (c_1, c_2) and let $\mathcal{E}^{TP}$ be the mean squared prediction errors over $\mathcal{J}$ for TP, i.e.

$$\begin{aligned}
\mathcal{E}^{TP} &= \frac{1}{T}\sum_{j=j_1}^{j_T}\left[\frac{1}{4n^2}\sum_{c_1=1}^{2n}\sum_{c_2=1}^{2n}\left(s^{TP}_{c_1,c_2} - u_{j,c_1,c_2}\right)^2\right] \\
&= \frac{1}{T}\sum_{j=j_1}^{j_T}\left[\, d_2\left(\mathbf{M}^{TP,j}, \mathbf{M}^{j}\right)\,\right] \\
&= \frac{1}{T}\sum_{j=j_1}^{j_T}\left[\sum_{p=1}^{K} f_{j,h_p}\sum_{q=1}^{K} f_{j,h_q}\left[\, p_{j,h_p,h_q}\left(s^{TP}_{h_p,h_q} - 1\right)^2\right.\right. \\
&\quad \left.\left. + \left(1 - p_{j,h_p,h_q}\right)\left(s^{TP}_{h_p,h_q} - 0\right)^2\right]\right]
\end{aligned}$$

Note that the expression of the normalized squared euclidean distance, d_2, in terms of frequencies and proportions is analogous to that of d_1 in (1).

Indeed f_{j,h_p}, f_{j,h_q} and p_{j,h_p,h_q} at locus j are defined as in (1). Estimates for $\left(s^{TP}_{i,h_p,h_q}\right)_{(p,q)\in\{1,..,K\}^2}$ are obtained by differentiating $\mathcal{E}^{TP}$ with respect to $s^{TP}_{h_p,h_q}$, i.e.

$$\frac{\partial \mathcal{E}^{TP}}{\partial s^{TP}_{h_p,h_q}} = 0 \Longleftrightarrow \hat{s}^{TP}_{h_p,h_q} = \frac{\sum_{j=j_1}^{j_T} f_{j,h_p} f_{j,h_q} p_{j,h_p,h_q}}{\sum_{j=j_1}^{j_T} f_{j,h_p} f_{j,h_q}}$$

Note that the second derivative of $\mathcal{E}^{TP}$ with respect to $s^{TP}_{h_p,h_q}$ is positive since it is a sum of frequencies. This implies that $\mathcal{E}^{TP}$ reaches a minimum for the set of estimates $\left(\hat{s}^{TP}_{h_p,h_q}\right)_{(p,q)\in\{1,..,K\}^2}$, since $\mathcal{E}^{TP}$ is a sum of convex functions of each $s^{TP}_{h_p,h_q}$. Hence, TP associates $\hat{s}^{TP}_{h_p,h_q}$ to any observed couple $\left(h_p, h_q\right)$ at any tested position $i \in \mathcal{I}$. The observed target SNPs ($j \in \mathcal{J}$) are used to estimate the predictions of allelic identity for TP and should not be confused with the unobserved tested positions ($i \in \mathcal{I}$).

Statistical models, test statistic and relative efficiency

Mixed models

The following mixed models were used to test for the presence of a QTL at a given position $i \in \mathcal{I}$ for all AIP:

$$\begin{cases} \mathbf{Y} = \mathbf{X}\beta + \mathbf{Z_h}\mathbf{h} + \mathbf{Z_u}\mathbf{u} + \boldsymbol{\varepsilon} & (\mathrm{H}_1) \\ \mathbf{Y} = \mathbf{X}\beta + \mathbf{Z_u}\mathbf{u} + \boldsymbol{\varepsilon} & (\mathrm{H}_0) \end{cases}$$

where β is a fixed effect, which is the overall mean, and $\mathbf{X} = \mathbb{1}_{\boldsymbol{n}}$ is a vector of n ones. Vector $\mathbf{u}$ represents the random polygenic effects due to relationships among individuals, i.e. $\mathbf{u} \sim \mathcal{N}_n\left(0, \mathbf{A}\sigma^2_{\mathrm{u}}\right)$ where $\mathbf{A}$ is the additive relationship matrix built from the pedigree [28,29]. $\mathbf{Z_h}$ and $\mathbf{Z_u}$ are design matrices that link random effects to individuals and ε is the vector of homoscedastic error terms, i.e. $\varepsilon \sim \mathcal{N}_n\left(0, \mathbf{I}_n\sigma^2_{\varepsilon}\right)$.

In the model corresponding to (H_1), $\mathbf{h}$ is a vector of random effects of haplotypes at position i, i.e. $\mathbf{h} \sim \mathcal{N}_\kappa(0, \mathbf{H}^{\mathcal{P},i}\sigma^2_{\mathrm{h}})$, where κ $(\kappa \leq K)$ corresponds to the number of observed haplotypes, or alleles, at position i. Note that $\mathbf{h}$ has the same dimension κ for all AIP except for $\mathrm{IBS_m}$ and Beagle. The tested positions coincide with marker positions for these two predictors. At a tested marker i, $\kappa = 2$ for $\mathrm{IBS_m}$ and κ is equal to the number of local clusters for Beagle. Therefore, depending on the predictor $\mathcal{P}$, $\mathbf{H}^{\mathcal{P},i}$ is a similarity matrix based on either distinct observed haplotypes (e.g. $\mathcal{P}$ = Score) or distinct clusters (e.g. $\mathcal{P}$ = Beagle). Note that $\mathbf{H}^{\mathcal{P},i}$ and $\mathbf{M}^{\mathcal{P},i}$ are equivalent sources of data contingent upon the list of haplotypes, or distinct local clusters, for the chromosome segments at any tested position. Indeed, depending on $\mathcal{P}$, one can build $\mathbf{M}^{\mathcal{P},i}$ from $\mathbf{H}^{\mathcal{P},i}$ in one of the two following ways. (1) $\mathbf{M}^{\mathcal{P},i}_{c_1,c_2} = \mathbf{H}^{\mathcal{P},i}_{h(c_1),h(c_2)}$, where $h(c_1)$ and $h(c_2)$ are the haplotype numbers carried by chromosomes c_1 and c_2, respectively or (2) $\mathbf{M}^{\mathcal{P},i}_{c_1,c_2} = \mathbf{H}^{\mathcal{P},i}_{\mathcal{C}(c_1),\mathcal{C}(c_2)}$, where $\mathcal{C}(c_1)$ and $\mathcal{C}(c_2)$ are the cluster numbers to which chromosomes c_1 and c_2 belong, respectively.

RLRT statistic

The Expectation-Maximization algorithm was used for the restricted maximum likelihoods of the mixed models [30-33], to estimate the components β, $\mathbf{h}$, $\mathbf{u}$, ε and the variance terms $\sigma^2_{\mathrm{h}}, \sigma^2_{\mathrm{u}}, \sigma^2_{\varepsilon}$. Let $\lambda^{\mathcal{P}}_i$ be the restricted maximum likelihood ratio test (RLRT) of (H_1) versus (H_0) for position i, i.e.

$$\lambda^{\mathcal{P}}_i = -2ln\left(\frac{L_{REML}(\mathrm{H}_0)}{L^{\mathcal{P}}_{REML}(\mathrm{H}_1)}\right)$$

We defined $\theta^{\mathcal{P}}_{\mathrm{m.a.}}$ as the estimated position of a QTL for a predictor $\mathcal{P}$, i.e.

$$\theta^{\mathcal{P}}_{\mathrm{m.a.}} = \underset{i\in\mathcal{I}}{argmax}\left\{\hat{\lambda}^{\mathcal{P}}_i\right\}$$

Relative efficiency

To compare the predictive ability of the different predictors in relation to LD, we defined $\theta^{\mathcal{P}}_{\mathrm{r.e.}}$ as the tested position where $d_1\left(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}\right)$ is minimized for a predictor $\mathcal{P}$, i.e.

$$\theta^{\mathcal{P}}_{\mathrm{r.e.}} = \underset{i\in\mathcal{I}}{argmin}\left\{d_1\left(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}\right)\right\}$$

Consequently, we defined the relative efficiency of a predictor $\mathcal{P}$ as follows. Predictor $\mathcal{P}$ is considered to be more efficient than a predictor $\mathcal{P}'$ if

$$\begin{cases} |\theta^{\mathcal{P}}_{\mathrm{r.e.}} - \theta_{QTL}| < |\theta^{\mathcal{P}'}_{\mathrm{r.e.}} - \theta_{QTL}| & (a) \\ d_1\left(\mathbf{M}^{\mathcal{P},\theta^{\mathcal{P}}_{\mathrm{r.e.}}}, \mathbf{M}^{QTL}\right) < d_1\left(\mathbf{M}^{\mathcal{P}',\theta^{\mathcal{P}'}_{\mathrm{r.e.}}}, \mathbf{M}^{QTL}\right) & (b) \end{cases}$$

where $|.|$ is the absolute value. When $\theta^{\mathcal{P}}_{\mathrm{r.e.}}$ was not unique, the mean of the different argmins was retained as $\theta^{\mathcal{P}}_{\mathrm{r.e.}}$. Inequality (a) states that the tested position associated with the best prediction, of the allele identity at the QTL, is closer to the QTL for $\mathcal{P}$ than that for $\mathcal{P}'$. Inequality (b) states that the true allelic identity at the QTL is better predicted by $\mathcal{P}$ at $\theta^{\mathcal{P}}_{\mathrm{r.e.}}$ than by $\mathcal{P}'$ at $\theta^{\mathcal{P}'}_{\mathrm{r.e.}}$.

Comparison criteria

N simulations $(w = 1, .., N)$ were performed to evaluate the mapping accuracy and the relative efficiency of the different AIP in different situations.

Mapping accuracy

The mapping accuracy of the simulated QTL was evaluated for each AIP with the root mean square error (RMSE):

$$\mathrm{RMSE}^{\mathrm{m.a.}} = \sqrt{\frac{1}{N}\sum_{w=1}^{N}\left(\theta^{\mathcal{P},w}_{\mathrm{m.a.}} - \theta_{QTL}\right)^2}$$

Relative efficiency

The relative efficiency of each AIP was evaluated by considering the three following quantities:

$$\begin{cases} \mathrm{RMSE}^{\mathrm{r.e.}} = \sqrt{\frac{1}{N}\sum_{w=1}^{N}\left(\theta^{\mathcal{P},w}_{\mathrm{r.e.}} - \theta_{QTL}\right)^2} \\ \hat{\mathbb{E}}^{\mathrm{r.e.}} = \frac{1}{N}\sum_{w=1}^{N} d_1\left(\mathbf{M}^{\mathcal{P},\theta^{\mathcal{P},w}_{\mathrm{r.e.}}}, \mathbf{M}^{QTL,w}\right) \\ \hat{\sigma}^{\mathrm{r.e.}} = \sqrt{\frac{1}{N}\sum_{w=1}^{N}\left(d_1\left(\mathbf{M}^{\mathcal{P},\theta^{\mathcal{P},w}_{\mathrm{r.e.}}}, \mathbf{M}^{QTL,w}\right) - \hat{\mathbb{E}}^{\mathrm{r.e.}}\right)^2} \end{cases}$$

where $\mathrm{RMSE}^{\mathrm{r.e.}}$ and $\hat{\mathbb{E}}^{\mathrm{r.e.}}$ measure conditions (a) and (b), defined in the paragraph on relative efficiency, and $\hat{\sigma}^{\mathrm{r.e.}}$ measures the standard deviation of the matrix distance at $\theta^{\mathcal{P}}_{\mathrm{r.e.}}$.

Data for simulation

A sliding window of $t = 6$ markers was chosen for all analyses, except for $\mathrm{IBS_m}$ and Beagle. Windows of six and 12 markers were previously shown to be optimal for QTL mapping accuracy [34,35] with 60K type

SNP chips. Hence, all analyses were done using a sliding window of $t = 6$ markers, except for IBS_m and Beagle, to make comparison between the series of results easier. A set of 90 human chromosomes 21 from unrelated Han Chinese individuals from Beijing (HCB), and a set of 235 swine chromosomes 18 from French Large White (FLW) pigs, were used for LD and matrix distance computations. The 90 HCB chromosomes were genotyped for 16 881 SNPs and are available from the HapMap project website (http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/2005-03_phaseI/full/). The FLW chromosomes were genotyped for 1252 SNPs using the Illumina Porcine 60K+SNP iSelect Beadchip [36]. Only 14 976 SNPs on the HCB chromosomes and 969 SNPs on the FLW chromosomes for which the minor allele frequency was greater than 5% were retained for analysis. The LD and matrix distance computations were conducted for the HCB and the FLW chromosomes. The QTL simulations were only conducted for the FLW chromosomes for which a pedigree was available. The marker density varied across the FLW chromosomes based on physical distance in kilobase. One megabase was considered equivalent to 1 cM for conversion in this study.

Variation of LD between tested positions and a QTL

LD between a tested position i and a QTL was measured using the multiallelic measure R of LD as suggested by [37-39]. Let $\mathbf{\Delta_p} = f^{QTL}_{i,h_p a_1} - f_{i,h_p} f_{a_1}$ be the LD coefficient between haplotype h_p at position i and allele a_1 at the QTL. $f^{QTL}_{i,h_p a_1}$ is the frequency of haplotype $h_p a_1$ defined by the marker haplotype h_p that spans position i and allele a_1 at the QTL. f_{a_1} is the frequency of allele a_1 at the QTL and f_{i,h_p} is haplotype h_p frequency at i. Note that $-\mathbf{\Delta_p} = f^{QTL}_{i,h_p a_2} - f_{i,h_p} f_{a_2}$. Hence, for a biallelic QTL, R can be expressed as:

$$R_{i,QTL} = \frac{\sum_{p=1}^{K}\sum_{l=1}^{2}\left(f^{QTL}_{i,h_p a_l} - f_{i,h_p} f_{a_l}\right)^2}{\left(1-\sum_{p=1}^{K} {f^{QTL}_{i,h_p}}^2\right)\left(1-\sum_{l=1}^{2} {f_{a_l}}^2\right)}$$

$$= \frac{\sum_{p=1}^{K}\left[\left(\mathbf{\Delta_p}\right)^2 + \left(-\mathbf{\Delta_p}\right)^2\right]}{\left(1-\sum_{p=1}^{K} {f^{QTL}_{i,h_p}}^2\right)\left(1-\sum_{l=1}^{2} {f_{a_l}}^2\right)}$$

$$= \frac{2\sum_{p=1}^{K} \mathbf{\Delta_p^2}}{\left(1-\sum_{p=1}^{K} {f^{QTL}_{i,h_p}}^2\right)\left(1-\sum_{l=1}^{2} {f_{a_l}}^2\right)} = \frac{D^2_{i,QTL}}{H_i . H_{QTL}}$$

where $H_i = 1 - \sum_{p=1}^{K} {f^{QTL}_{i,h_p}}^2$ and $H_{QTL} = 1 - \sum_{l=1}^{2} {f_{a_l}}^2$ are the Hardy-Weinberg heterozygosities at i and the QTL respectively and $D^2_{i,QTL} = 2\sum_{p=1}^{K} \mathbf{\Delta_p^2}$. $R_{i,QTL}$ and $D^2_{i,QTL}$ are expected to increase as the tested position i gets closer to a QTL. The general behaviors of the normalized measure $R_{i,QTL}$ and the non-normalized measure $D^2_{i,QTL}$ were described by computing the LD between the haplotypes at successive distinct positions, using a sliding window, and the alleles of a fixed SNP centered over a region of 81 markers on the chromosomes. The fixed SNP was centered over a region of 76 distinct overlapping sliding windows available within the region of 81 markers. The 76 distinct positions associated to the windows played the role of the tested positions of an association study. The fixed SNP played the role of a biallelic QTL. The computation was repeated for all possible regions of 81 successive markers. Since 969 SNPs were retained on the 235 porcine chromosomes, computation was performed for 889 ($969 - 81 + 1 = 889$) regions of 81 markers. The same procedure was performed on the HCB chromosomes, thus leading to 14 896 possible regions for this set of chromosomes. The empirical means of the 889 FLW and the 14 896 HCB LD profiles were then computed to describe the expected behaviors of $R_{i,QTL}$ and $D^2_{i,QTL}$. Another major purpose of these computations was to help the analytical comparison of the AIP and the associated matrix distances, which can be expressed as elements of multiallelic LD (see Results section).

Distributions of matrix distance as a function of multiallelic LD

The distributions of the matrix distance for the six compared AIP, as function of local multiallelic LD, were also evaluated on the FLW and HCB chromosomes. The matrix distances for the six AIP were calculated at 966 and 14 973 possible target SNPs for the FLW and HCB chromosomes, respectively. The target SNPs were defined in exactly the same way as used for the trained predictor (TP). The matrix distances calculated at each window that harbors a target SNP for the six AIP were then plotted against the multiallelic measure R of LD between the haplotypes and the target alleles within the window.

QTL simulation on FLW chromosomes

The 235 FLW chromosomes were included in $N = 200$ gene-drop simulations, in a 25-generation pedigree for the FLW breed, using the LDSO software [40]. The pedigree was composed of 1594 founders, 3373 sires and 7100 dams. The gene-drop procedure was used to generate different realistic genealogy structures between the chromosomes. For each gene-drop the 235 FLW chromosomes were uniformly distributed, with replacement, among the 1594 founders of the pedigree. Hence, the measured LD structure for mapping among descendant individuals at the end of each gene-drop was almost the same as on the 235 FLW chromosomes. It must be emphasized that the use of replicates of only 235 chromosomes to populate 1594 diploid founders, followed by 25 generations

of recombinations events, means that the number of different haplotypes at a position is much lower than 3188 (2×1594). Thus, the results correspond to medium range population sizes. After each gene-drop, only the chromosomes and phenotypes of the $n = 485$ individuals of generation 25 were retained for subsequent analyzes.

Three distant SNPs were chosen as putative QTL, in order to have different LD levels with the six-marker haplotype that surrounds them on the 235 initial FLW chromosomes. Two different QTL effects were simulated for each of these SNPs, thus leading to six different scenarios. The LD between these SNPs and the observed haplotypes that harbored them was measured using the multiallelic measure R of LD. The LD levels around the three SNPs were equal to 0.52, 0.18 and 0.08, and the lengths of the haplotypes harboring them were equal to 0.09 cM, 0.37 cM and 0.75 cM, respectively. Note that these differences in length were due to the different marker densities in the distinct regions that harbor each putative QTL. The length of the region scanned for QTL mapping around each simulated QTL was approximately 3 cM.

The phenotypes in the pedigree were computed as $y_i = \frac{1}{2}\left(p_i^f + p_i^m\right) + \phi_i + g_i^{QTL} + \delta$, where p_i^f, p_i^m are normal random polygenic effects of the parents with variance 0.5, ϕ_i is a normal random mendelian sampling effect with variance 0.25 and δ is a normal random environmental effect with variance 1. g_i^{QTL} is the QTL genotype effect of individual i. QTL genotype effect was first computed as $g_i^{QTL} = 2$ or 0 or -2, if the QTL genotype of individual i was a_1a_1 or a_1a_2 or a_2a_2 respectively. In the same way a second set of simulations was carried out with the QTL genotype effect computed as $g_i^{QTL} = 0.5$ or 0 or -0.5. Only the gene-drop simulations for which the minor allele frequency at the QTL was greater or equal to 0.1 were retained. Each simulated QTL was verified for Hardy-Weinberg equilibrium during simulations. Hence, under the standard model, where the dominance effect is equal to 0 as in this study, the first simulated QTL effect explained at most 57% of the phenotypic variance for equal frequencies at the QTL. In the same way, the second simulated QTL effect explained at most 8% of the phenotypic variance.

Results

This section gives theoretical and empirical results that show that, compared to others, some AIP exhibit a better behavior for the decrease of their matrix distance, as defined by expression (1), when the multiallelic LD between a tested position and a QTL increases. In summary, the theoretical results show that expression (1) can be written as a function of the multiallelic LD coefficients of R, and that the decreasing behavior of this function depends on the nature of the AIP (see equations (2), (3), (4), (5) and (6) of this section). The empirical results show that R is expected to be highest when the tested position is closest to the QTL (see Figure 2 of this section). The expectation taken for the multiallelic LD was the empirical mean, which was found to converge for distant regions on the chromosomes. These regions can be assumed to be independent, thus showing an expected behavior for the multiallelic LD. The empirical results also show that the tested position that minimizes the matrix distance is highly correlated with the mapping accuracy of the AIP (see sub-section on mapping accuracy and relative efficiency of this section).

Variation of LD between tested positions and a QTL

Figure 2 shows the empirical means of the 889 FLW and the 14 896 HCB LD profiles for $R_{i,QTL}$ and $D^2_{i,QTL}$.

In Figure 2 the values of $R_{i,QTL}$ and $D^2_{i,QTL}$ increase, as expected, as the tested position i moves closer to the QTL. This implies that the sum of the $\mathbf{\Delta_p^2}$ terms increases on average as position i moves toward the QTL. The highest expected values for $R_{i,QTL}$ and $D^2_{i,QTL}$, in Figure 2 are reached for the tested position closest to the QTL. Note that the range of values for $D^2_{i,QTL}$, in Figure 2 is smaller than that of $R_{i,QTL}$. This is due to the lack of a normalization factor for $D^2_{i,QTL}$.

Matrix distance as function of multiallelic LD coefficients

Based on expression (1), $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ can be re-written as [see Additional file 1]:

$$\begin{aligned} d_1\left(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}\right) = \sum_{p=1}^{K}\sum_{q=1}^{K}\Big[\Big[f^{QTL}_{i,h_pa_1}f^{QTL}_{i,h_qa_1} \\ + f^{QTL}_{i,h_pa_2}f^{QTL}_{i,h_qa_2}\Big]\left(1 - s^{\mathcal{P}}_{i,h_p,h_q}\right) \\ + \left[f^{QTL}_{i,h_pa_2}f^{QTL}_{i,h_qa_1} + f^{QTL}_{i,h_pa_1}f^{QTL}_{i,h_qa_2}\right]s^{\mathcal{P}}_{i,h_p,h_q}\Big] \end{aligned} \quad (2)$$

Replacing the $2K$ frequencies in expression (2) by the $(\mathbf{\Delta_p})_{1\le p\le K}$ LD coefficient terms and the product of marginal frequencies gives [see Additional file 1]:

$$\begin{aligned} d_1\left(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}\right) &= \sum_{p=1}^{K}\Bigg[4\left(\sum_{q\neq p}^{K} s^{\mathcal{P}}_{i,h_p,h_q} - s^{\mathcal{P}}_{i,h_p,h_p}\right)\mathbf{\Delta_p^2} \\ &\quad + \Psi^{\mathcal{P}}_{pq}\left(\Delta_{l\neq p,q}\right)\mathbf{\Delta_p} + \Phi^{\mathcal{P}}_{pq}\left(\Delta_{l\neq p,q}\right)\Bigg] \\ &= \xi^{\mathcal{P}}\left(\mathbf{\Delta_1}, .., \mathbf{\Delta_K}\right), \end{aligned} \quad (3)$$

where $\Psi^{\mathcal{P}}_{pq}\left(\Delta_{l\neq p,q}\right)$ and $\Phi^{\mathcal{P}}_{pq}\left(\Delta_{l\neq p,q}\right)$ are sums and products of marginal frequencies, allelic identity predictions and LD coefficient terms. The general behavior

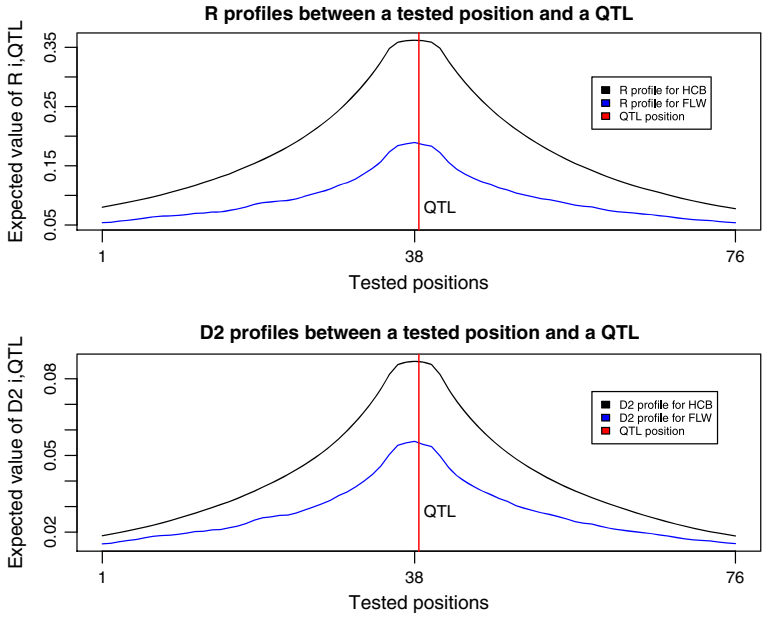

Figure 2 Empirical means of the 889 FLW and 14 896 HCB LD profiles, obtained for the normalized and the non-normalized multiallelic LD between tested positions (tested position *i* = center of six marker haplotypes) and a biallelic QTL (red vertical line), for regions of 81 markers on chromosomes.

of $\xi^{\mathcal{P}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$, with respect to $R_{i,QTL}$, is unspecifiable due to its complexity. For instance, the behavior of $\xi^{\mathcal{P}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$ cannot be specified for continuous AIP in [0,1]. However for $\mathcal{P} = \text{IBS}_{\text{hap}}$, $\xi^{\mathcal{P}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$ reduces to a sum of strictly concave functions of each LD coefficient [see Additional file 1], i.e.

$$\begin{aligned}\xi^{\text{IBS}_{\text{hap}}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K}) &= \sum_{p=1}^{K}\left[-4\Delta_{\mathbf{p}}^{\mathbf{2}} + \Psi_{pq}^{\text{IBS}_{\text{hap}}}\boldsymbol{\Delta_p} + \Phi_{pq}^{\text{IBS}_{\text{hap}}}\right] \\ &= \sum_{p=1}^{K}\left[Q_p(\boldsymbol{\Delta_p})\right], \end{aligned} \quad (4)$$

where $\Psi_{pq}^{\text{IBS}_{\text{hap}}}$ and Φ_{pq} are sums and products of marginal frequencies that do not depend on LD coefficients. Let $\boldsymbol{\Delta_p^*} = \frac{\Psi_{pq}^{\text{IBS}_{\text{hap}}}}{8}$ be the critical value for each Q_p function. $\xi^{\text{IBS}_{\text{hap}}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$ will decrease if the squared or absolute deviation of each $\boldsymbol{\Delta_p}$ term from its corresponding $\boldsymbol{\Delta_p^*}$ increases [see Additional file 2]. However note that the squared deviations of all $\boldsymbol{\Delta_p}$ terms from their corresponding critical values do not need to increase simultaneously for $\xi^{\text{IBS}_{\text{hap}}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$ to decrease. For example, some Q_p functions corresponding to haplotypes with low frequencies can be negligible in expression (4). Hence, if $\sum_{p=1}^{K}\left(\boldsymbol{\Delta_p} - \boldsymbol{\Delta_p^*}\right)^2$ increases sufficiently, $\xi^{\text{IBS}_{\text{hap}}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$ will decrease. It can be shown that $\sum_{p=1}^{K}\left(\boldsymbol{\Delta_p} - \boldsymbol{\Delta_p^*}\right)^2$ will increase if $\sum_{p=1}^{K}\boldsymbol{\Delta_p}^2$ increases and that these two quantities share almost the same pattern for their expected values [see Additional file 2]. Thus, according to the $D_{i,QTL}^2$ profiles in Figure 2, $\xi^{\text{IBS}_{\text{hap}}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$ is expected to decrease as position i moves toward the QTL position.

An important result for $\xi^{\mathcal{P}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$ is obtained when only two haplotypes are observed among the K possible haplotypes. In this case, $\xi^{\mathcal{P}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$ reduces to a real function of one LD coefficient [see Additional file 1], i.e.:

$$\begin{aligned}\xi^{\mathcal{P}}(\boldsymbol{\Delta_1}) = &\left[-4s_{i,h_1,h_1}^{\mathcal{P}} - 4s_{i,h_2,h_2}^{\mathcal{P}} + 8s_{i,h_1,h_2}^{\mathcal{P}}\right]\boldsymbol{\Delta_1^2} \\ &+ \Psi^{\mathcal{P}}\boldsymbol{\Delta_1} + \Phi^{\mathcal{P}}, \end{aligned} \quad (5)$$

where $\Psi^{\mathcal{P}}$ and $\Phi^{\mathcal{P}}$ are terms independent of LD, and the minimum and maximum possible values for $\boldsymbol{\Delta_1}$ are given by $-\frac{1}{4}$ and $\frac{1}{4}$, respectively. If $\mathcal{P} = \text{IBS}_{\text{hap}}$ we have:

$$\xi^{\text{IBS}_{\text{hap}}}(\boldsymbol{\Delta_1}) = -8\boldsymbol{\Delta_1^2} + \Psi^{\text{IBS}_{\text{hap}}}\boldsymbol{\Delta_1} + \Phi^{\text{IBS}_{\text{hap}}} \quad (6)$$

The minimum and maximum possible values for the critical value, $\boldsymbol{\Delta_1^*}$, of $\xi^{\text{IBS}_{\text{hap}}}$ are given by $-\frac{1}{4}$ and $\frac{1}{4}$, respectively, if the tested locus and the QTL are monomorphic [see Additional file 1]. In other words, the

critical value of this function will always lie within the range of the LD coefficient when LD exist. In expression (5), the coefficient $\left[-4s^{\mathcal{P}}_{i,h_1,h_1} - 4s^{\mathcal{P}}_{i,h_2,h_2} + 8s^{\mathcal{P}}_{i,h_1,h_2}\right]$ is always greater or equal to -8 for any other predictor $\mathcal{P}$ than IBS_{hap}, since $s^{\mathcal{P}}_{i,h_1,h_2} \in [0,1]$. For instance, AIP by construction assign positive values to $s^{\mathcal{P}}_{i,h_1,h_2}$ when haplotypes h_1 and h_2 share allele similarity. This property is even truer if h_1 and h_2 are very similar. In such cases, the highest rate of decrease for $\xi^{\mathcal{P}}$, with respect to the absolute deviation of $\mathbf{\Delta_1}$ from $\mathbf{\Delta^*_1}$, is thus induced by $\mathcal{P} = \text{IBS}_{\text{hap}}$. Moreover, for such cases, we also have $\xi^{\mathcal{P}}\left(-\frac{1}{4}\right) = \xi^{\mathcal{P}}\left(\frac{1}{4}\right) \in \left[\frac{1}{2}s^{\mathcal{P}}_{i,h_1,h_2}, 1\right]$, which expresses a lower bound for $\xi^{\mathcal{P}}$ (i.e. $\frac{1}{2}s^{\mathcal{P}}_{i,h_1,h_2}$, [see Additional file 1]). Finally, $\xi^{\mathcal{P}}\left(-\frac{1}{4}\right) = \xi^{\mathcal{P}}\left(\frac{1}{4}\right) = 0$ if and only if $\mathcal{P} = \text{IBS}_{\text{hap}}$. In other words, when LD between the haplotypes and the QTL alleles is complete, the matrix distance is equal to 0 if and only if $\mathcal{P} = \text{IBS}_{\text{hap}}$ [see Additional file 1]. The decreasing behavior of $\xi^{\mathcal{P}}$ between a tested position and a QTL for a substantial increase of LD is therefore deteriorated for AIP with continuous predictions in $[0, 1]$. Hence, this result questions the behavior of AIP with continuous predictions in $[0, 1]$ in relation to LD, in the general case where K is greater than 2.

Distributions of matrix distance as function of multiallelic LD

Figures 3 and 4 show the distributions of the matrix distance for the six AIP against the local multiallelic LD. Figures 3 and 4 convey only local information for the case where the tested position is closest to the QTL, as opposed to Figure 2. Darker and lighter blue regions in Figures 3 and 4 correspond to higher and lower density of points. The red lines in Figures 3 and 4 correspond to non-parametric LOESS regressions of the matrix distance on the multiallelic LD.

Figures 3 and 4 show a better behavior of IBS_{hap} and P(IBD) for the decrease of their matrix distance, with lower variability around the LOESS curves compared to the other predictors, when the LD between the haplotypes and the target alleles increases. The distributions of the matrix distance for IBS_{hap} and P(IBD) in these figures show similar trends on the FLW and HCB chromosomes. This is due to the fact that these two predictors perform similarly in some conditions (see sub-section on mapping accuracy and relative efficiency). However IBS_{hap} shows a better behavior compared to all other predictors in Figures 3 and 4, for the decrease of its matrix distance with increasing multiallelic LD. The good behavior of IBS_{hap} for the decrease of its matrix distance in Figures 3 and 4 is totally explained by equation (4), where the sum of the concave polynomials decreases as the multiallelic LD increases. The better behavior of IBS_{hap}, compared to the other predictors in Figures 3 and 4, is explained by equations (3) and (5), which show that continuous predictions in [0,1] will deteriorate the decrease of the matrix distance with respect to LD.

The matrix distances for Beagle and IBS_{m} were also plotted against the local multiallelic LD between the haplotypes and the target alleles in Figures 3 and 4, although these two predictors are defined for marker positions only. Indeed, one of the aims of this study was to compare the AIP based on local LD between haplotypes and alleles at a hidden locus. TP, Score, Beagle and IBS_{m} showed poor relationships for the decrease of their matrix distance with the increasing multiallelic LD. The matrix distance distributions showed high variability for these predictors with respect to R on the FLW and HCB chromosomes. Note that the length of the six marker haplotypes on the HCB chromosomes were equal to 0.01 cM, on average, compared to 0.31 cM on average for those on the FLW chromosomes.

Mapping accuracy and relative efficiency

Table 1 relates the relative efficiency of the six AIP that were compared, and their mapping accuracies, for a QTL simulated under six scenarios on the FLW chromosomes for $N = 200$ simulations. $R_{i^*,QTL}$ in Table 1 corresponds to the multiallelic LD at position i^*, measured between the marker-haplotypes that harbor the simulated QTL and the QTL alleles. Note that the tested position i^* does not necessarily coincide with the QTL position. Thus, i^* can be defined as the tested position closest to the simulated QTL.

In Table 1, $\text{IBS}^{\text{QTL}}_{\text{m}}$ refers to the IBS_{m} predictor applied to the data set containing the causal variants. This situation was examined as a gold standard. As shown in Table 1 and as expected, $\text{IBS}^{\text{QTL}}_{\text{m}}$ provided the best mapping accuracy since the data set used contained the causal variants and both the simulated QTL and the analyzed markers were biallelic. However, it should be noted that the $\text{RMSE}^{\text{m.a.}}$ for $\text{IBS}^{\text{QTL}}_{\text{m}}$ was never equal to 0. This is principally due to the error term in the probabilistic models for hypothesis testing. $\text{RMSE}^{\text{r.e.}}$ for $\text{IBS}^{\text{QTL}}_{\text{m}}$ was also not equal to 0 when LD was highest ($R_{i^*,QTL} = 0.52$). This was due to a nearby marker which was in complete LD with the SNP that simulated the QTL (i.e. the biallelic LD was complete). Consequently the argmin for the set of matrix distances was not unique.

In Table 1 both $\text{RMSE}^{\text{r.e.}}$ and $\text{RMSE}^{\text{m.a.}}$ increased globally for all predictors when LD decreased in the vicinity of the QTL. $\text{RMSE}^{\text{r.e.}}$ and $\text{RMSE}^{\text{m.a.}}$ were highly correlated, regardless of the QTL effect. Across all LD levels, the Spearman correlation coefficient between these two quantities was equal to 0.89 (or 0.91) when the QTL

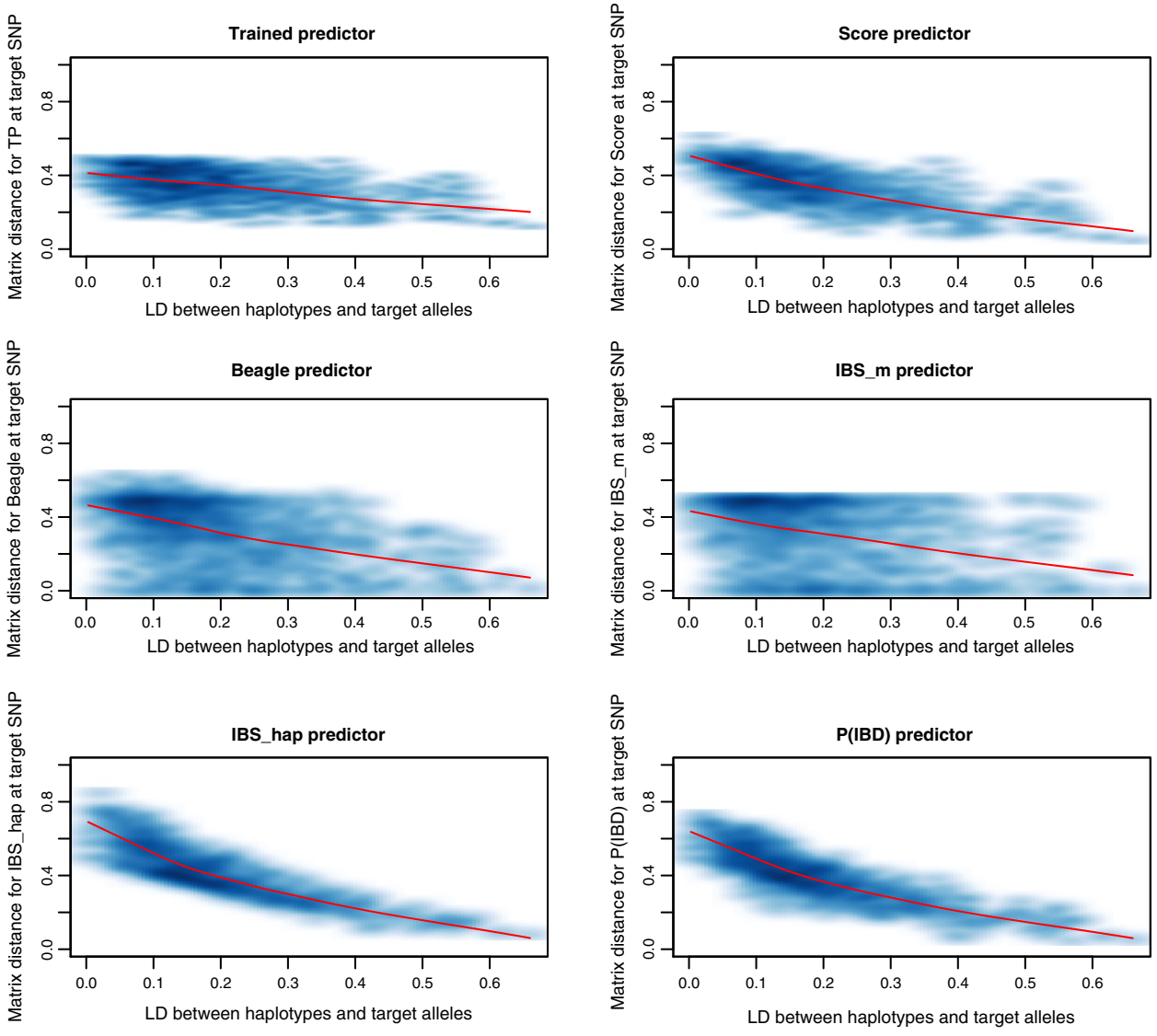

Figure 3 The matrix distance at 966 target SNPs on the 235 FLW chromosomes for the six compared AIP. Darker and lighter blue regions correspond to higher and lower density of points. The red lines correspond to LOESS regressions of the matrix distance on the multiallelic LD.

effect explained at most 57% (or 8%) of the total variance, respectively (Figure 5).

Each dot in Figure 5 represents $RMSE^{m.a.}$ against $RMSE^{r.e.}$ for one of the AIP at a particular LD level. In Table 1, the IBS_{hap} predictor was often the most accurate and efficient AIP when the data was analyzed without the QTL. However, the P(IBD) predictor showed similar mapping and efficiency results to IBS_{hap}. As defined by [1], the P(IBD) predictor relies on the IBS state of alleles between haplotype markers which suggests that IBS_{hap} and P(IBD) may perform similarly in some conditions [41]. Indeed, the distribution of IBD probabilities in the vicinity of a simulated QTL was almost bimodal (0 or 1) among the different pairs of chromosome segments for the different sets of simulations, and thus similar to the distribution of the values for IBS_{hap} between the segments. To illustrate this phenomenon, Figure 6 provides an example of distributions for the values of P(IBD) and IBS_{hap}, for one gene-drop simulation, between pairs of chromosome segments around the simulated QTL for the moderate LD situation ($R_{i^*,QTL} = 0.18$).

IBS_{hap} and P(IBD) also showed similar patterns at a set of tested positions for the matrix distances $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$. Figure 7 shows an example for the mean and the sample quantiles at 2.5 and 97.5% for $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ at each tested position for the six AIP, from 200 gene-drop simulations with a QTL simulated for the moderate LD situation ($R_{i^*,QTL} = 0.18$).

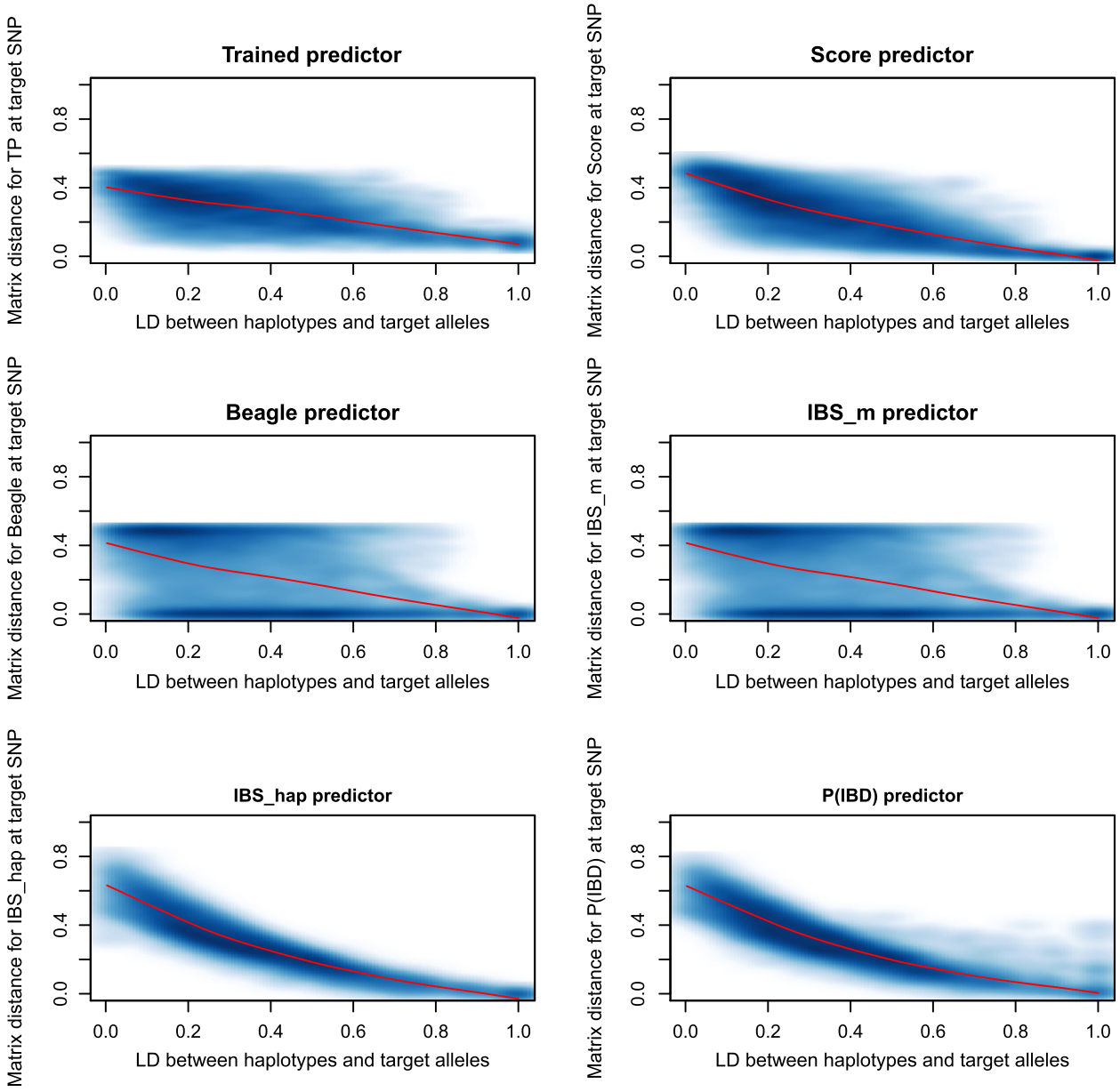

Figure 4 The matrix distance at 14973 target SNPs on the 90 HCB chromosomes for the six compared AIP. Darker and lighter blue regions correspond to higher and lower density of points. The red lines correspond to LOESS regressions of the matrix distance on the multiallelic LD.

As observed in Figure 7, the minima of the curves for the mean and the sample quantiles at 2.5 and 97.5% of the matrix distance distributions almost coincide with the QTL position for IBS_{hap} and P(IBD). For these two predictors, the three curves also show a smooth decreasing behavior as the tested position gets closer to the simulated QTL. This behavior shows the ability of IBS_{hap} and P(IBD) to capture LD structure along the chromosomes with respect to the simulated QTL, for different gene-drop simulations. It is interesting to note that IBS_{hap} and P(IBD) show similar patterns for the mean and the sample quantiles curves. However, the minimum of each of the three curves in Figure 7 is lower for IBS_{hap} than for P(IBD). Note that the patterns of the matrix distances for IBS_{hap} in Figure 7 are explained by equation (4) and Figure 2. That is, the matrix distance will decrease for IBS_{hap} due to the expected increase of the multiallelic LD, as the tested position moves toward the QTL position. In the same way, the patterns of the matrix distances for P(IBD) in Figure 7 are explained according to Figures 2 and 6. That is, P(IBD)

Table 1 Relative efficiencies and mapping accuracies for different QTL effects

	AIP			IBS_m^{QTL}	IBS_m	TP	Score	IBS_{hap}	P(IBD)	Beagle
$R_{i^*,QTL} = 0.52$	Relative efficiency		$RMSE^{r.e.}$	0.02	0.16	0.17	0.15	0.06	0.10	0.54
			$\hat{\mathbb{E}}^{r.e.}$	0.03	0.01	0.23	0.14	0.12	0.14	0.28
			$\hat{\sigma}^{r.e.}$	0.09	0.02	0.02	0.03	0.04	0.03	0.04
	Mapping	$\leq 57\%$	$RMSE^{m.a.}$	0.11	0.17	0.23	0.17	0.10	0.11	0.16
	accuracy	$\leq 8\%$	$RMSE^{m.a.}$	0.17	0.22	0.32	0.45	0.28	0.26	0.40
$R_{i^*,QTL} = 0.18$	Relative efficiency		$RMSE^{r.e.}$	0.00	0.18	0.46	0.21	0.14	0.14	0.40
			$\hat{\mathbb{E}}^{r.e.}$	0.00	0.18	0.39	0.35	0.31	0.34	0.31
			$\hat{\sigma}^{r.e.}$	0.00	0.06	0.02	0.03	0.05	0.04	0.06
	Mapping	$\leq 57\%$	$RMSE^{m.a.}$	0.06	0.29	0.27	0.33	0.16	0.21	0.28
	accuracy	$\leq 8\%$	$RMSE^{m.a.}$	0.10	0.34	0.36	0.46	0.29	0.30	0.31
$R_{i^*,QTL} = 0.08$	Relative efficiency		$RMSE^{r.e.}$	0.00	0.76	1.00	1.00	1.04	1.04	0.72
			$\hat{\mathbb{E}}^{r.e.}$	0.00	0.24	0.35	0.33	0.31	0.37	0.34
			$\hat{\sigma}^{r.e.}$	0.00	0.06	0.04	0.05	0.06	0.05	0.08
	Mapping	$\leq 57\%$	$RMSE^{m.a.}$	0.13	0.66	0.58	0.54	0.51	0.58	0.55
	accuracy	$\leq 8\%$	$RMSE^{m.a.}$	0.18	0.71	0.66	0.71	0.62	0.69	0.59

- $R_{i^*,QTL}$: Multiallelic measure of LD between the simulated QTL and the haplotypes harboring it.
- $RMSE^{r.e.}$: Root mean square error of $\theta_{r.e.}^{\mathcal{P}}$ with respect to θ_{QTL} (cM).
- $\hat{\mathbb{E}}^{r.e.}$: Expected value of the matrix distance at $\theta_{r.e.}^{\mathcal{P}}$.
- $\hat{\sigma}^{r.e.}$: Standard error of the matrix distance at $\theta_{r.e.}^{\mathcal{P}}$.
- $RMSE^{m.a.}$: Root mean square error of $\theta_{m.a.}^{\mathcal{P}}$ with respect to θ_{QTL} (cM).

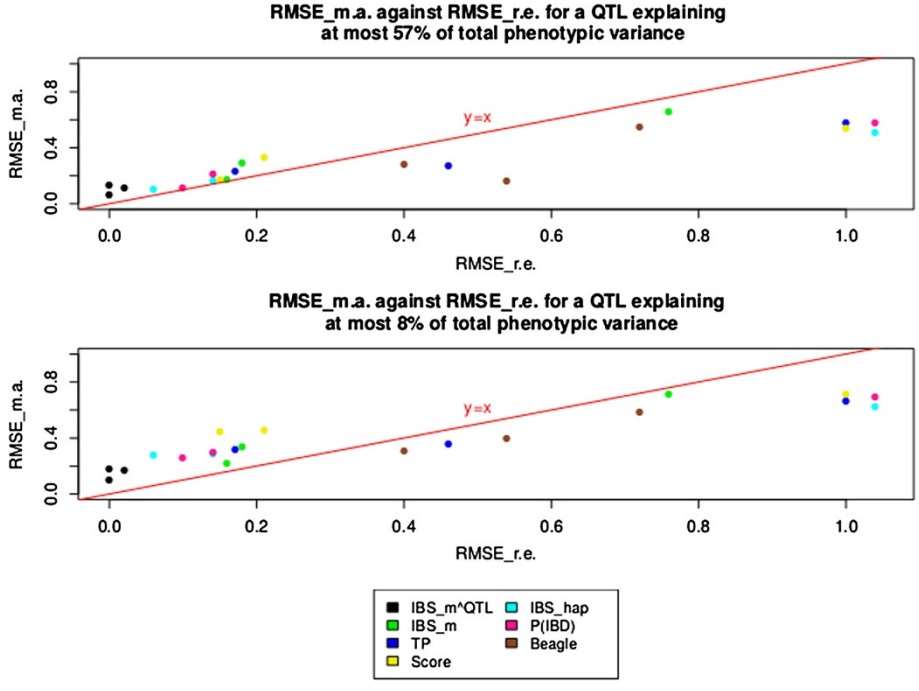

Figure 5 $RMSE^{m.a.}$ against $RMSE^{r.e.}$ for the compared AIP according to different QTL effects across all LD levels.

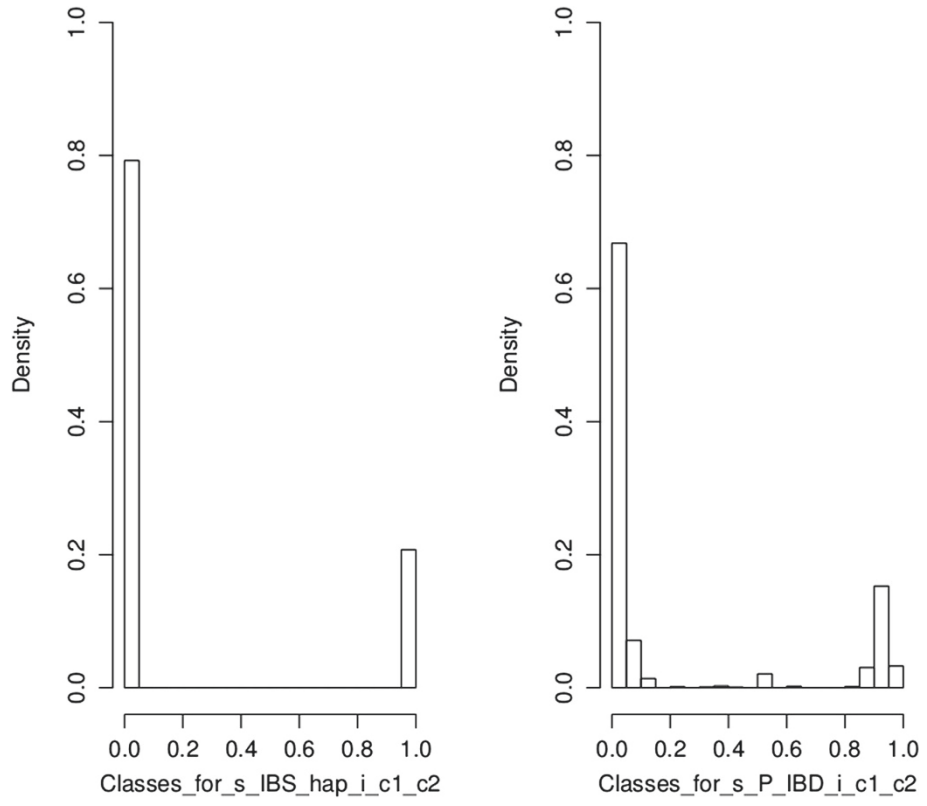

Figure 6 Distribution of values for IBS_{hap} and P(IBD) between chromosome segments around the simulated QTL for the moderate LD situation ($R_{i^*,QTL} = 0.18$), example for one simulation. The class width for the IBD probabilities is equal to 0.05.

will behave slightly differently from IBS_{hap}, according to Figures 2 and 6, when taking equations (3) and (5) into account.

As shown in Figure 7, the other predictors cannot capture the LD structure along the chromosomes with respect to the simulated QTL as well as IBS_{hap} and P(IBD); this is particularly the case for Score and even more for TP. For the latter two predictors, $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ shows little variability and is low on average across the tested positions. This could explain the lack of a clear ranking between the mapping accuracies of TP and Score in Table 1. For Beagle, a good relative efficiency and mapping accuracy was observed for the lowest LD situation ($R_{i^*,QTL} = 0.08$) in Table 1, compared to all the other predictors, when the QTL effect was low. Note that AIP that are based on haplotypes and that do not perform haplotype clustering like Beagle, may not be at an advantage for a low LD situation. For example, the matrix distance for IBS_{hap}, as defined by equation (4), will not decrease if there is little LD between local haplotypes and QTL alleles. Therefore, haplotype clustering is necessary for such situations. Moreover, these AIP will intrinsically provide an excess of degrees of freedom for testing association if the QTL is biallelic, while not compensating for the low LD captured in the matrix distance. Hence, AIP based on haplotype clustering can provide higher mapping accuracy for low LD situations.

Discussion

Matrix distance properties

The present study showed that the QTL mapping accuracy of AIP is highly correlated to the tested position that minimizes the matrix distance defined for comparison. The use of the matrix distance to compare various AIP has many advantages for methodology development and validation. First, it is independent of phenotype simulation processes and statistical tests that are commonly used to compare QTL mapping accuracy of different AIP [4,8,23,25]. Indeed the phenotype simulation process, when based on certain specific assumptions, may favor some AIP over others: for example, IBD-based AIP might be at an advantage if the phenotypes are simulated only according to population history. The statistical test used may also favor some AIP, such as IBS_{hap}, IBS_{m} and Beagle, over others due to numerical stability when estimating variance components. As such, solving mixed model equations when covariance matrices are close to singularity due to AIP computation has been reported as an

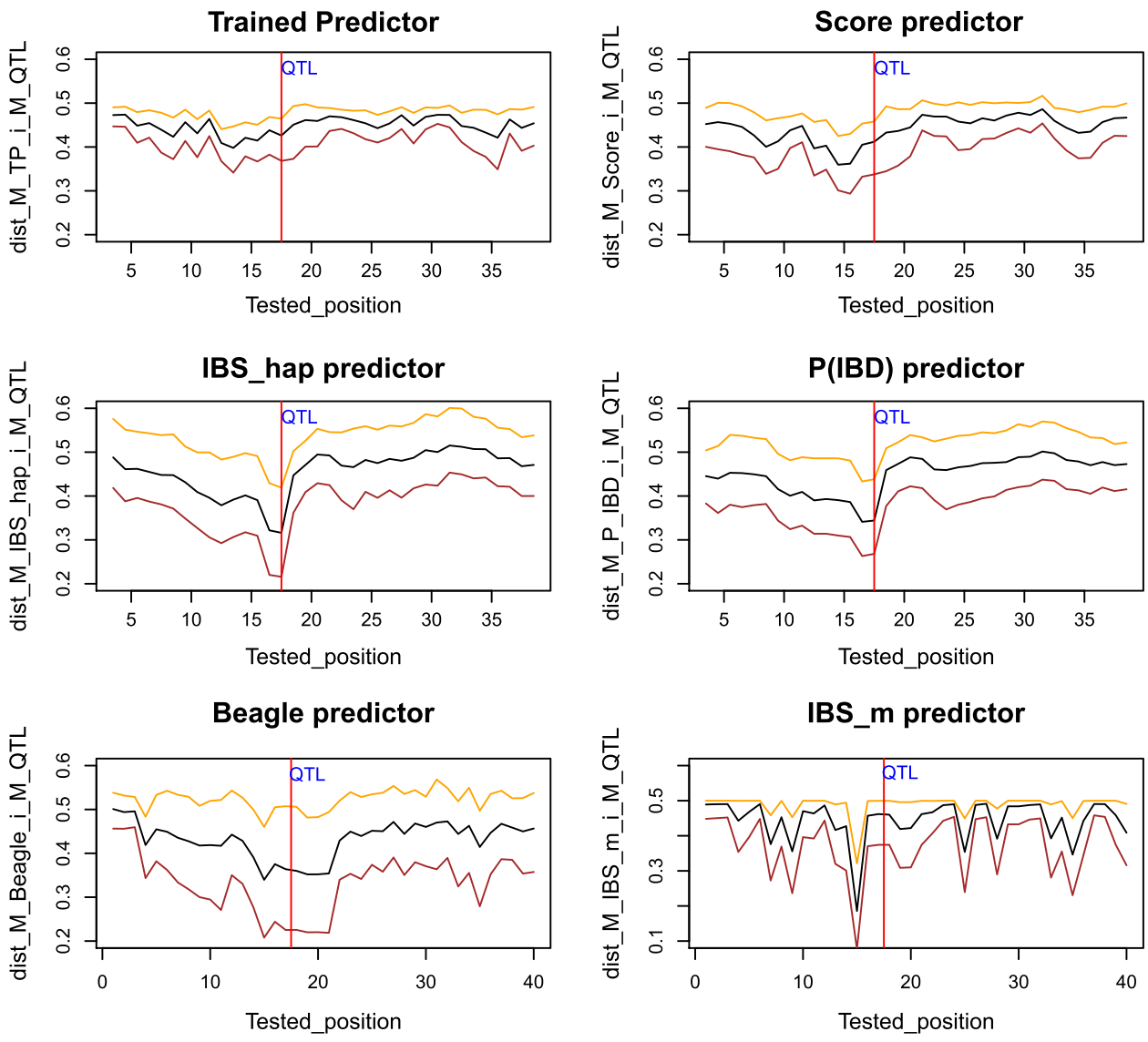

Figure 7 The mean and sample quantiles at 2.5% and 97.5% of $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ (black, brown and orange respectively) for TP, Score, IBS_{hap}, P(IBD), Beagle and IBS_m, with respect to the tested position and the simulated QTL position (red line) for the moderate LD situation ($R_{i^*,QTL} = 0.18$).

issue, and clustering strategies for haplotypes, which actually modify the properties of the AIP matrices, have been proposed to facilitate computation [42,43]. The major drawback of the matrix distance approach is related to this advantage: a particularly efficient AIP or a particularly efficient haplotype size, identified from the matrix distance, that can not be used in association studies would be of no value. Another advantage of the matrix distance approach is that computation time is highly reduced compared to association studies, so numerous comparisons can be done. In the present study, the relative efficiency of the AIP was consistent with the results for QTL mapping accuracy, regardless of the QTL effects and LD patterns. Therefore, the concept of relative efficiency was proven useful to compare AIP and avoid time-consuming association studies on simulated data. Combining the relative efficiency with the mapping accuracy of predictors could also be helpful to gain a better understanding of the underlying mechanisms in an association study.

Comparing AIP

The results showed that the most accurate AIP for mapping were those that best captured LD between a tested position and a QTL. This was proposed from algebraic developments in the simplest situations and validated

using real data and simulations. The matrix distance can be written for any AIP as a sum of functions of LD coefficients, and more precisely for the IBS_{hap} predictor as a sum of concave polynomials of LD coefficients. When LD was moderate to high around the QTL, the IBS_{hap} predictor was the most efficient and accurate matrix for mapping. For a biallelic QTL, the domains of values for which some of these concave polynomials can either decrease or increase with increasing LD was shown in our developments as limited to extreme allele frequencies for the haplotypes and QTL. Additionally, continuous AIP in [0,1] were shown to deteriorate the matrix distance generally when LD between a tested position and the QTL increased. This was observed on two unrelated data sets, which showed that this behavior is not related to the marker density or population history. All LD measures are based on counting occurrences for discrete events at distinct loci to quantify non-random association [37,38], which thus explains the algebraic and simulation results for discrete and continuous AIP when a relatively high LD is available for detection. The pig example was built using 235 haplotypes and 25 gene-drop generations, a realistic situation with regard to the effective population size. However, the impact of the resulting long-range haplotypic identity, which depends strongly on the population size and mating strategies, on the relative values of the considered AIP should be investigated.

Despite using two contrasting data sets in terms of marker density and population history, P(IBD) always behaved very similarly to IBS_{hap}. When extending the calculations to longer haplotypes (results not shown), a similar behavior was observed. Yet advantages have been reported for P(IBD) compared to IBS_{hap} in some situations. For example, Roldan *et al.* [43] showed better accuracy for P(IBD) compared to IBS_{hap}, after a clustering step for haplotypes when marker intervals were equal to 0.05 cM between SNPs, but not when they reached 0.25 cM. However in Roldan *et al.* [43], different statistical models were applied to P(IBD) versus IBS_{hap} (mixed model *versus* fixed effects model respectively). Hence, these two AIP were not compared on the same basis. For instance, Boleckova *et al.* [44] showed that statistical models in which haplotypes were fitted as random effects performed better than those in which they were fitted as fixed effects. When both LD and the QTL effect were low, Beagle showed a relatively good efficiency and mapping accuracy. It was not possible to derive algebraical comparisons between AIP when LD was low, but this, together with earlier studies that point out that continuous advanced methods are more efficient than simple IBS_{hap}, suggests that some continuous AIP in [0,1] may provide efficiency when LD between markers and a QTL is reduced.

Extending the results to multiallelic QTL

In the present study, we considered a biallelic QTL for algebra and simulations. Yet the algebraic derivation of the matrix distance can be generalized to a multiallelic QTL without difficulty [see Additional file 1]. As suggested by these developments, for a multiallelic QTL, the relationship between continuous predictions of allelic identity at a tested position and the corresponding LD coefficients will tend to be looser than for discrete predictions. In addition, the matrix distance for the IBS_{hap} predictor can always be written as a sum of concave polynomials of LD coefficients for any degree of allelism at the QTL.

Conclusion

The IBS_{hap} predictor can always capture multiallelic LD between a tested position and a QTL, regardless of the degree of allelism at the QTL. The IBS_{hap} predictor also has the advantage of being simple, fast and numerically stable when used in association studies. Therefore, it is suggested that, for studies with a high density of markers and for which LD between markers and the causal variants is likely to be high, the use of the IBS_{hap} predictor is recommended.

Additional files

Additional file 1: Algebraic derivations of formulas in the main text. This file contains all the algebraic derivations for expressions (1) to (6) and the generalization of the matrix distance, as a sum of concave functions of LD coefficients when $\boldsymbol{\mathcal{P}} = IBS_{hap}$, for the case of a multiallelic QTL.

Additional file 2: Domains of LD coefficients and boundary conditions for the critical values of each Q_p function. This file contains the domain of values for the multiallelic LD coefficients, the boundary conditions for the critical value of each Q_p function in expression (4) and the relation between the sum of the squared deviations and $D^2_{I,QTL}$.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LJ derived the analytical results, performed the simulations and wrote the manuscript. LJ, JME and HG were involved in the conception of the study. All authors read and approved the final manuscript.

Acknowledgements

The study was supported by the French National Research Agency (ANR-09-GENM-002 Rules & Tools Project).

Author details

[1]INRA, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), F-31326, Castanet-Tolosan, France. [2]Université de Toulouse, INP, ENSAT, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), F-31326, Castanet-Tolosan, France. [3]Université de Toulouse, INP, ENVT, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), F-31076, Toulouse, France.

Received: 20 November 2013 Accepted: 20 May 2014
Published: 14 July 2014

References

1. Meuwissen THE, Goddard ME: **Prediction of identity by descent probabilities from marker haplotypes.** *Genet Sel Evol* 2001, **33:**605–634.

2. Li J, Jiang T: **Haplotype-based linkage disequilibrium mapping via direct data mining.** *Bioinformatics* 2005, **21:**4384–4393.
3. Browning SR: **Multilocus association mapping using variable-length Markov chains.** *Am J Hum Genet* 2006, **78:**903–913.
4. Pong-Wong R, George AW, Woolliams JA, Haley CS: **A simple and rapid method for calculating identity-by-descent matrices using multiple markers.** *Genet Sel Evol* 2001, **33:**453–471.
5. Bercovici S, Meek C, Wexler Y, Geiger D: **Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping.** *Bioinformatics* 2010, **26:**i175–i182.
6. Druet T, Georges M: **A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping.** *Genetics* 2010, **184:**789–798.
7. Akey J, Jin L: **Xiong M: Haplotypes vs single marker linkage disequilibrium tests: what do we gain?** *Eur J Hum Genet* 2001, **9:**291–300.
8. Abdallah J, Goffinet B, Cierco-Ayrolles C, Pérez-Enciso M: **Linkage disequilibrium fine mapping of quantitative trait loci: a simulation study.** *Genet Sel Evol* 2003, **35:**513–532.
9. Cardon LR, Abecasis GR: **Using haplotype blocks to map human complex trait loci.** *Trends Genet* 2003, **19:**136–140.
10. Clark AG: **The role of haplotypes in candidate gene studies.** *Genet Epidemiol* 2004, **27:**321–333.
11. Schaid DJ: **Evaluating associations of haplotypes with traits.** *Genet Epidemiol* 2004, **27:**348–364.
12. Browning BL, Browning SR: **Efficient multilocus association testing for whole genome association studies using localized haplotype clustering.** *Genet Epidemiol* 2007, **31:**365–375.
13. Chen Y, Li X, Li J: **A novel approach for haplotype-based association analyzis using family data.** *BMC Bioinformatics* 2010, **11:**S45.
14. Lin WY, Yi N, Zhi D, Zhang K, Gao G, Tiwari HK, Liu N: **Haplotype-based methods for detecting uncommon causal variants with common SNPs.** *Genet Epidemiol* 2012, **36:**572–582.
15. Knüppel S, Esparza-Gordillo J, Marenholz I, Holzhütter HG, Bauerfeind A, Ruether A, Weidinger S, Lee YA, Rohde K: **Multi-locus stepwise regression: a haplotype based algorithm for finding genetic associations applied to atopic dermatitis.** *BMC Med Genet* 2012, **13:**8.
16. Li M, Wing HW, Art BO: **A sparse transmission disequilibrium test for haplotypes based on Bradley-Terry graphs.** *Hum Hered* 2012, **73:**52–61.
17. Risch N, Merikangas K: **The future of genetic studies of complex human diseases.** *Science* 1996, **273:**1516–1517.
18. Terwilliger JD, Weiss KM: **Linkage disequilibrium mapping of complex disease: fantasy or reality?** *Curr Opin Biotechnol* 1998, **9:**578–594.
19. Jorde LB: **Linkage disequilibrium and the search for complex disease genes.** *Genome Res* 2000, **10:**1435–1444.
20. Weiss KM, Clark AG: **Linkage disequilibrium and the mapping of complex human traits.** *Trends in Genet* 2002, **18:**19–24.
21. Slatkin M: **Disequilibrium mapping of a quantitative-trait locus in an expanding population.** *Am J Hum Genet* 1999, **64:**1764–1772.
22. Farnir F, Coppieters W, Arranz JJ, Berzi P, Cambisano N, Grisart B, Karim L, Marcq F, Moreau L, Mni M, Nezer C, Simon P, Vanmanshoven P, Wagenaar D, George M: **Extensive genome-wide linkage disequilibrium in cattle.** *Genome Res* 2000, **10:**220–227.
23. Meuwissen THE, Goddard ME: **Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci.** *Genetics* 2000, **155:**421–430.
24. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES: **Structure of linkage disequilibrium and phenotypic associations in the maize genome.** *Proc Natl Acad Sci USA* 2001, **98:**11479–11484.
25. He W, Fernando RL, Dekkers JCM, Gilbert H: **A gene frequency model for QTL mapping using Bayesian inference.** *Genet Sel Evol* 2010, **42:**21.
26. Grapes L, Dekkers JCM, Rothschild MF, Fernando RL: **Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci.** *Genetics* 2004, **166:**1561–1570.
27. Browning BL, Browning SR: **Haplotypic analysis of Wellcome Trust Case Control Consortium data.** *Human Genet* 2008, **123:**273–280.
28. Henderson CR: **Best linear unbiased estimation and prediction under a selection model.** *Biometrics* 1975, **31:**423–447.
29. Henderson CR: **A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values.** *Biometrics* 1976, **32:**69–83.
30. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from in-complete data via the EM algorithm.** *Roy Statist Soc Ser B* 1977, **39:**1–38.
31. Patterson HD, Thompson R: **Recovery of inter-block information when block sizes are unequal.** *Biometrika* 1971, **58:**545–554.
32. Harville DA: **Bayesian inference for variance components using only error contrasts.** *Biometrika* 1974, **61:**383–385.
33. Foulley JL: **EM algorithm: theory and application to the mixed model.** *J Soc Fr Stat* 2002, **143:**57–109.
34. Calus MPL, Meuwissen THE, Windig JJ, Knol EF, Schrooten C, Vereijken ALJ, Veerkamp RF: **Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values.** *Genet Sel Evol* 2009, **41:**11.
35. Grapes L, Firat MZ, Dekkers JCM, Rothschild MF, Fernando RL: **Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci using identity by descent.** *Genetics* 2005, **172:**1955–1965.
36. Ramos AM, Crooijmans RP, Affara NA, Amaral AJ, Archibald AL, Beever JE, Bendixen C, Churcher C, Clark R, Dehais P, Hansen MS, Hedegaard J, Hu ZL, Kerstens HH, Law AS, Megens HJ, Milan D, Nonneman DJ, Rohrer GA, Rothschild MF, Smith TP, Schnabel RD, Van Tassell CP, Taylor JF, Wiedmann RT, Schook LB, Groenen MA: **Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology.** *PLoS ONE* 2009, **4:**e6524.
37. Hedrick PW, Thomson G: **A two-locus neutrality test: applications to humans, E. coli and Lodgepole pine.** *Genetics* 1985, **112:**135–156.
38. Hedrick PW: **Gametic disequilibrium measures: proceed with caution.** *Genetics* 1987, **117:**331–341.
39. Maurer HP, Knaak C, Melchinger AE, Ouzunova M, Frisch M: **Linkage disequilibrium between SSR markers in six pools of elite lines of an european breeding program for hybrid maize.** *Maydica* 2006, **51:**269–279.
40. Ytournel F, Teyssèdre S, Roldan D, Erbe M, Simianer H, Boichard D, Gilbert H, Druet T, Legarra A: **LDSO: A program to simulate pedigrees and molecular information under various evolutionary forces.** *J Anim Breed Genet* 2012, **129:**417–421.
41. Ytournel F, Gilbert H, Boichard D: **Concordance between IBD probabilities and linkage disequilibrium.** In *Proceedings of European Federation of Animal Science Annual Meeting; 26 August 2007; Dublin*; 2007:1248. [http://www.eaap.org/Previous_Annual_Meetings/2007Dublin/Papers/S38_1248_Ytournel.pdf]
42. Druet T, Fritz S, Boussaha M, Ben-Jemaa S, Guillaume F, Derbala D, Zelenika D, Lechner D, Charon C, Boichard D, Gut I, Eggen A, Gautier M: **Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map.** *Genetics* 2008, **178:**2227–2235.
43. Roldan DL, Gilbert H, Henshall JM, Legarra A, Elsen JM: **Fine-mapping quantitative trait loci with a medium density marker panel: efficiency of population structures and comparison of linkage disequilibrium linkage analysis models.** *Genet Res Camb* 2012, **94:**223–234.
44. Boleckova J, Christensen OF, Sørensen P, Sahana G: **Strategies for haplotype-based association mapping in a complex pedigreed population.** *Czech J Anim Sci* 2012, **1:**1–9.

doi:10.1186/1297-9686-46-45
Cite this article as: Jacquin *et al.*: Using haplotypes for the prediction of allelic identity to fine-map QTL: characterization and properties. *Genetics Selection Evolution* 2014 **46**:45.