

A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping

Tom Druet¹ and Michel Georges

Unit of Animal Genomics, GIGA-Research and Department of Animal Production, Faculty of Veterinary Medicine, University of Liège, B-4000 Liège, Belgium

Manuscript received August 11, 2009

Accepted for publication December 5, 2009

ABSTRACT

Faithful reconstruction of haplotypes from diploid marker data (phasing) is important for many kinds of genetic analyses, including mapping of trait loci, prediction of genomic breeding values, and identification of signatures of selection. In human genetics, phasing most often exploits population information (linkage disequilibrium), while in animal genetics the primary source of information is familial (Mendelian segregation and linkage). We herein develop and evaluate a method that simultaneously exploits both sources of information. It builds on hidden Markov models that were initially developed to exploit population information only. We demonstrate that the approach improves the accuracy of allele phasing as well as imputation of missing genotypes. Reconstructed haplotypes are assigned to hidden states that are shown to correspond to clusters of genealogically related chromosomes. We show that these cluster states can directly be used to fine map QTL. The method is computationally effective at handling large data sets based on high-density SNP panels.

ARRAY technology now allows genotyping of large cohorts for thousands to millions of single nucleotide polymorphisms (SNPs), which are becoming available for a growing list of organisms including human and domestic animals. Among other applications, these advances permit systematic scanning of the genome to map trait loci by association (*e.g.*, WELLCOME TRUST CASE CONTROL CONSORTIUM 2007; CHARLIER *et al.* 2008), to predict genomic breeding values for complex traits (MEUWISSEN *et al.* 2001; GODDARD and HAYES 2009), or to identify signatures of selection (*e.g.*, VOIGHT *et al.* 2006).

Present-day genotyping platforms do not directly provide information about linkage phase; *i.e.*, co-inherited alleles at adjacent heterozygous markers (haplotypes) are not identified as such. As haplotype information may considerably empower genetic analyses, indirect phasing strategies have been devised: haplotypes can be reconstructed from unphased genotypes using either familial information (Mendelian segregation and linkage) and/or population information (linkage disequilibrium, LD, and surrogate parents) (*e.g.*, WINDIG and MEUWISSEN 2004; SCHEET and STEPHENS 2006; KONG *et al.* 2008).

Haplotype-based approaches are routinely applied in animal genetics for combined linkage and LD mapping

of QTL (*e.g.*, MEUWISSEN and GODDARD 2000; BLOTT *et al.* 2003). In these studies, phasing has so far relied on familial information provided by the extended pedigrees typical of livestock (*e.g.*, WINDIG and MEUWISSEN 2004). This approach, however, leaves a nonnegligible proportion of genotypes unphased, especially for the less connected individuals. After phasing, identity-by-descent (IBD) probabilities conditional on haplotype data—needed for QTL mapping—are computed for all chromosome pairs, using familial as well as population information (hence combined linkage and LD mapping – L + LD) (*e.g.*, MEUWISSEN and GODDARD 2001). However, the use of high-density SNP chips and the analysis of ever larger cohorts render the computation of pairwise IBD probabilities a bottleneck.

We herein propose a more efficient, heuristic approach based on hidden Markov models (HMM). It simultaneously phases and sorts haplotypes in clusters that can be used directly for mapping or other purposes. The proposed method exploits familial as well as population information, and imputes missing genotypes. We herein describe the accuracy of the proposed method and its use for L + LD mapping of QTL.

MATERIALS AND METHODS

Haplotype reconstruction and clustering: Due to systematic recording of familial relationship in domestic animals, individuals are usually part of extended pedigrees. For the purpose of haplotype reconstruction we first split the pedigree such as

Supporting information is available online at <http://www.genetics.org/cgi/content/full/genetics.109.108431/DC1>.

¹Corresponding author: Unit of Animal Genomics, GIGA-R B34, 1 avenue de l'Hôpital, B-4000 Liège, Belgium. E-mail: tom.druet@ulg.ac.be

to consider only relationships between genotyped parents and genotyped offspring. In the ensuing subpedigrees, individuals can be parent only, offspring only, or parent and offspring.

Homozygous SNPs are considered phased *de facto*. The phasing of heterozygous genotypes proceeds in three steps:

Step I: Mendelian segregation. Marker alleles of heterozygous offspring are assigned to the paternal or maternal homolog following Mendelian segregation rules. In offspring, this leaves only SNPs unphased for which typed parents have the same heterozygous genotype as the offspring.

Step II: Linkage.

- a. Parents: Parental phases are completed on the basis of allelic cosegregation in the offspring following DRUET *et al.* (2008). This process requires heterozygous “anchoring” markers whose alternate alleles define the padumnal *vs.* madumnal homolog of the parent. For individuals that are both parent and offspring, step I will have defined several such anchoring markers. For parents without genotyped parents, it is impossible to determine the parental origin of the homologs. In this situation, we identify the most informative marker on the basis of the number of offspring for which the transmitted allele can be determined without ambiguity and design it as the “primordial anchoring” marker: one of its alleles defines the “left” homolog (inherited from parent A) and the other the “right” homolog (inherited from parent B). We then sequentially phase remaining markers with respect to already anchored ones as described in step 3 in DRUET *et al.* (2008). In brief, the probability that marker allele 1 belongs to the “left” (L) (respectively “right,” R) homolog is computed as $P^{L(R)} = L^{L(R)} / (L^L + L^R)$. In this, $L^{L(R)}$ corresponds to the likelihood that the marker allele belongs to the L (respectively R) haplotype of the parent, conditional on information from flanking anchoring markers. $L^{L(R)}$ is computed as $L_{UP}^{L(R)} \times L_{DN}^{L(R)}$, in which $L_{UP(DN)}^{L(R)}$ correspond to the likelihood conditional on information from anchoring markers located upstream (UP) (respectively downstream, DN) of the marker to be phased. $L_{UP(DN)}^{L(R)}$ are computed as

$$\prod_{i=1}^x (1 - r_i) \prod_{j=1}^y r_j,$$

where x (respectively y) is the number of offspring with marker phase in agreement (respectively disagreement) with the tested haplotype configuration of the examined parent and r_i (r_j) is the recombination rate between the tested marker and the nearest (respectively upstream or downstream) anchoring marker informative in offspring i (respectively j). For subpedigrees spanning more than two generations, parents are treated by increasing generation number, *i.e.*, starting with the oldest generation. Step IIa is repeated until no new marker is phased.

- b. Offspring: Heterozygous markers that remain unphased in offspring can be further treated, conditional on the known parental phase (determined in steps I and IIa), according to step 4 in DRUET *et al.* (2008). Using information from flanking markers phased in both parent and offspring, the probability that a marker allele belongs to the R (respectively L) homolog of the offspring can be computed as a function of intermarker recombination rates. A marker is considered to be phased if one of the two possible configurations has a probability exceeding a chosen threshold (0.999 in this study).

Linkage information was extracted only for parents with six or more offspring (IIa), and offspring with five or more sibs

(IIb), as it was found to inflate error rates for smaller pedigrees.

Steps I and II yield partially phased genotypes: These can be subdivided into “base” or “founder” haplotypes inherited from ungenotyped parents (assigned “left” *vs.* “right” status), and “descendent” haplotypes inherited from genotyped parents (assigned “padumnal” *vs.* “madumnal” status).

Step III: Linkage disequilibrium: To complete haplotype reconstruction, we exploit LD information using algorithms developed either in *fastPHASE* (SCHEET and STEPHENS 2006) or in *Beagle* (S. R. BROWNING and B. L. BROWNING 2007), with some modifications. Both approaches rely on HMM (RABINER 1989).

In the *fastPHASE* probability model (FPM), observed haplotypes are modeled as mosaics of K hidden states (HS) (or ancestral haplotypes), with K held constant throughout the genome. Parameters of the HMM are (i) population frequencies of each HS, which may differ between marker positions, (ii) HS-specific allele frequencies at each marker position, and (iii) “recombination rates” for each marker interval. The population frequencies of the HS at the first marker position define the initial-state probabilities. Transition probabilities are computed as a function (see below) of the recombination rate and the population frequencies of the different HS at the next marker position. The HS-specific allele frequencies define the observation (or emission) probabilities. As more than one allele may have nonzero frequency at a given marker position, HS in effect define clusters of related haplotypes.

The probability of a given Markov chain can be computed as

$$\pi_k^1 \theta_{kl}^1 \prod_{m=2}^M a_{kk'}^m \theta_{k'l}^m, \quad (1)$$

in which π_k^1 is the initial-state probability (*i.e.*, the probability that the Markov chain starts in HS k), θ_{kl}^m is the emission probability (the probability to observe allele l at marker position m in HS k), $a_{kk'}^m$ is the transition probability (the probability to move from HS k at marker m to HS k' at marker $m + 1$). $a_{kk'}^m$ equals $J_m \pi_{k'}^{m+1}$ when $k \neq k'$ and $(1 - J_m) + J_m \pi_{k'}^{m+1}$ when $k = k'$, where J_m is the probability to have a jump (recombination) between markers m and $m + 1$ (*i.e.*, the probability to change HS between markers m and $m + 1$) and $\pi_{k'}^{m+1}$ is the probability that the Markov chain moves to HS k' when a recombination occurred between markers m and $m + 1$ (corresponding to the population frequency of HS k' at marker $m + 1$). Both J_m and $\pi_{k'}^{m+1}$ are independent of HS at marker m .

Different values of K (from 5 to 30 with steps of five) were tested (data not shown). Models performed better in terms of haplotyping or imputation with increasing K , albeit with diminishing returns. Since computing time was also increasing with larger K , a compromise value of 20 was used in the present study.

The *Beagle* probability model uses a localized haplotype clustering model (LHCM), which can be interpreted as a special class of HMM (S. R. BROWNING and B. L. BROWNING 2007). In this, the number K of HS is allowed to vary across the genome and is determined by the number of edges (at the corresponding marker position) of a directed acyclic graph (DAG) summarizing all haplotypes encountered in the population (S. R. BROWNING and B. L. BROWNING 2007). The DAG defines (i) the population frequencies of each HS at each marker position (obtained from edge counts), (ii) allelic frequencies for each HS at each marker position, and (iii) nonzero transition probabilities coinciding with sites of bifurcation in the DAG. Generating the DAG requires specification of node-joining criteria (BROWNING 2006) including the scale and shift parameters m and b , which were set to 1.0

and 0.0 in BROWNING (2006) and to 4.0 and 0.2 in B. L. BROWNING and S. R. BROWNING (2007). We tested (data not shown) both set of parameters plus an intermediate solution (2.0 and 0.1). As the latter performed best, it was selected for further analysis.

With the LHCM model, the probability of a given Markov chain is also computed according to Equation 1 with, however, some important differences: (i) J_m and $\pi_{k'}^{m+1}$ are no longer independent of HS at marker m , and values of $a_{kk'}^m$ are directly estimated from edge counts; and (ii) while the DAG imposes θ_{kl}^m values of either 1 or 0 (according to the allele labeling the corresponding edge), these were set at 0.999 or 0.001 to accommodate genotyping errors.

Both HMM were applied in two stages. In the first stage, the models were trained on a haploid training set consisting of the partially phased base haplotypes obtained after steps I and II. In the *fastPHASE* option this step generated EM parameter estimators. In the *Beagle* option it generated an optimal DAG. The latter was obtained iteratively: missing alleles in the base haplotypes were initially imputed randomly (*i.e.*, equal probability for the two alleles), and the resulting augmented haplotypes were used to build the first DAG. In subsequent iterations, missing alleles were updated by sampling from the probability distribution of HS. In the second stage, we completed the actual haplotype reconstruction and clustering by running a diploid HMM on the complete data set. The diploid HMM simultaneously models two independent chains, corresponding respectively to the “left” and “right” homolog of the individual. $K_L \times K_R$ HS combinations are considered at each marker position, where K_L and K_R are the number of HS states characterizing the “left” and “right” HMM, respectively. Assume that an individual has genotype 12 at marker position m . If the phase is known to be “1_{left}2_{right}” from steps I and/or II, the joint emission probability at that position equals $\theta_{k_1,1}^m \theta_{k_2,2}^m$. If the phase remains unknown after steps I and II, the joint emission probability equals $\theta_{k_1,1}^m \theta_{k_2,2}^m + \theta_{k_1,2}^m \theta_{k_2,1}^m$.

The diploid HMM was first applied on the base individuals, *i.e.*, genotyped individuals without genotyped parents. The same HMM ($K_L = K_R = K$) was used to model left and right homolog with EM parameters (FPM) or optimal DAG (LHCM) as estimated on the haploid training set. For each base individual, we determined the most likely HS composition of its constituent haplotypes using the Viterbi algorithm.

We subsequently treated the descendent individuals using a modified diploid HMM. In this, haplotypes derived from genotyped parents were modeled as a mosaic of the two previously established (steps I and II) haplotypes of the parent ($K = 2$) (Guo 1994), while residual base haplotypes (if any) were modeled using the full HMM. Thus, the number of HS combinations to consider at a given marker position is 2×2 for individuals with two descendent haplotypes, and $2 \times K$ for individuals with one descendent and one base haplotype. The latter “hybrid” situation is typical of paternal half-sib pedigrees with ungenotyped dams. The Markov chain for a descendent haplotype can still be written as

$$\pi_k^1 \theta_{kl}^1 \prod_{m=1}^{M-1} a_{kk'}^m \theta_{k'l}^m$$

with, however, the following changes: only two HS were allowed corresponding to the left and right homolog of the parent, π_k^1 was set at 0.50 for both homologs, θ_{kl}^m was set at 0.999 for the allele present in parental haplotype k and 0.001 for the other allele (to accommodate genotyping errors), and $a_{kk'}^m$ was set at $(1 - r_m)$ and r_m when $k = k'$ or $k \neq k'$, respectively, where r_m is the recombination rate between markers m and $m + 1$. The modified diploid HMM applied on descendent individuals in effect extracts linkage information only (offspring with

two genotyped parents) or linkage and LD information jointly (offspring with one genotyped parent). Finally, the “population-wide” HS status of the base haplotypes was projected on their descendent haplotypes.

Knowledge of HS status in base and descendent individuals allows (i) phasing of the markers that remained unresolved after I and II and (ii) imputation of genotypes at missing marker positions. The corresponding data augmentation was achieved by sampling unresolved phases and missing genotypes according to their respective probabilities computed from the allele-specific emission probabilities of the constituent HS.

The FPM was run with a home-made program (DualPHASE). Two programs were used to implement the LHCM. First, Beagle 2.1.3 (S. R. BROWNING and B. L. BROWNING 2007) was used to construct the optimal DAG from a set of base haplotypes. Then, DAGPHASE (developed as part of this work) reads the DAG and uses the derived HMM to sample haplotypes with a haploid model or to apply the Viterbi algorithm with the diploid model.

QTL mapping: Testing for the presence of a QTL at a given map position was performed using the following mixed model

$$y = Xb + Z_h h + Z_u u + e,$$

in which b is a vector of fixed effects, which in this study reduces to the overall mean, h is the vector of random QTL effects corresponding to the K defined HS, u is the vector of random individual polygenic effects (HENDERSON 1984), and e is the vector of individual error terms. HS effects with corresponding variance, σ_{H}^2 , individual polygenic effects with corresponding variance, σ_A^2 , and individual error terms with corresponding variance, σ_E^2 , were estimated by restricted maximum likelihood (REML) analysis (JOHNSON and THOMPSON 1995). The covariance between individual polygenic effects corresponded to twice the coefficient of coancestry (A -matrix computed from available genealogical information) times the additive genetic variance (σ_A^2). The covariance between the different HS effects was assumed to be zero, hence modeling QTL with a finite number of alleles. The presence of a QTL at a given map position was tested by a likelihood ratio test (LRT) (GEORGE *et al.* 2000).

This novel approach was compared with the standard L + LD QTL mapping approach relying on the computation of pairwise IBD probabilities according to MEUWISSEN and GODDARD (2001), followed by the clustering of chromosomes with IBD probabilities exceeding 0.90 according to DRUET *et al.* (2008). The haplotypes used in this analysis were the ones generated with DualPHASE.

Simulated pedigrees: We simulated SNP genotype data in a population structure typical of dairy cattle as follows. We first simulated 1000 haplotypes of length 100 Mb using the *ms* software (HUDSON 2002). In this, we assumed an effective population size of 5000, a mutation rate of 10^{-8} per base and generation, and a recombination rate of 1cM/Mb. We randomly selected 4% or 20% of the SNPs with minor allele frequency (MAF) higher than 0.05, yielding respectively ~ 2300 (23 SNPs/Mb) and $\sim 12,000$ (120 SNPs/Mb) markers closely mimicking genome-wide densities of $\sim 70,000$ and $\sim 360,000$ SNPs. Randomly selected haplotypes were then dropped in a real Dutch Holstein-Friesian pedigree composed of 212,162 animals spanning up to 20 generations. From this pedigree, we sampled 100 sires and their 969 sons, with complete pedigree information for at least six generations. The number of sons per sire ranged from 1 to 66. Sampled individuals were either parent or offspring but never both. The procedure was repeated 100 times, yet always based on the sampling of the same individuals in the pedigree.

Real pedigree: Real data consisted of a Dutch Holstein-Friesian pedigree composed of 199 sires and their 1502 sons, for a total of 1702 bulls (160/199 sires also appeared as sons). The number of sons per sires ranged from 1 to 32 (average 8.4). There were 80, 59, and 35 half-sib families with more than 5, 10, or 20 sons, respectively. All animals were genotyped for 2130 BTA14 SNP markers covering ~81 Mb. Haplotyping was performed with the complete data set, yet results summarized for two subregions with high and low marker density, respectively. The high-density subregion consisted of the first 500 markers (84 SNP/Mb) while the low-density subregion corresponded to the last 256 markers (12 SNP/Mb). Phenotypes were daughter yield deviations (DYD) for milk fat content (VANRADEN and WIGGANS 1991). The residual variance of each DYD was weighted by the corresponding number of daughter equivalents (VANRADEN and WIGGANS 1991).

RESULTS

Haplotyping: Haplotyping accuracy was evaluated by the number of “switches” per individual, *i.e.*, the number of erroneous inferences of linkage phase between successive heterozygous markers. It was estimated both for the simulated and real data, which were very similar in terms of MAF and level of LD as a function of distance (measured as r^2) (supporting information, Figure S1).

For the simulated data, the inferred phase was simply compared with the true phase and the number of switches counted. Numbers of SNPs phased, with corresponding numbers of switches, were compiled (a) after extraction of the familial information, (b) after addition of the population information, and (c) for the complete process. Figures were collated for individuals sorted according to the richness of the familial information (none, Mendelian I, linkage IIa, I + IIa, I + IIa + IIb). The results obtained can be summarized as follows (Table 1):

- i. DualPHASE and DAGPHASE performed remarkably well, even for individuals for whom no familial information was available or utilized (*i.e.*, sires with less than six offspring). For such individuals, switches affected approximately 1/900 (high-density map) and 1/200 (low-density map) intervals when using the LHCM model, or 1/150 and 1/100 intervals when using FPM, corresponding respectively to 4.0, 3.7, 24.9, and 8.2 switches for a 100-Mb-long chromosome.
- ii. These results can be improved even more by judiciously exploiting familial information when available. Extracting Mendelian information (step I) to phase offspring was always advantageous. It allows error-free phasing (assuming error-free genotyping) of half of the offsprings’ heterozygous markers. Subsequent phasing of the remaining heterozygous markers using population information was ameliorated, leading to overall switch rates of 1/1000 (high density) and 1/400 (low density) or 1/600 and 1/300 for LHCM and FPM, respectively.
- iii. Extracting type IIa linkage information to phase parental genotypes proved very valuable as long as the number of available offspring was sufficient. In the specific case of paternal half-sib pedigrees with ungenotyped dams, type IIa information extracted from less than six offspring led to a marked inflation in switch rate (data not shown). Above this threshold, however, more than 98% of the parent’s markers (irrespective of map density) could be phased with switch rates around 1/6000 (low density) and 1/35000 (high density). Importantly, improving the phase of a parent by exploiting type IIa information had a positive effect on the quality of phase reconstruction of its offspring when using that information in step III. Indeed, the switch rate in the offspring was thereby reduced to less than 1/6000 (low density) and 1/10000 (high density).
- iv. Extracting type IIb linkage information to phase offspring prior to the application of step III never proved beneficial. It always introduced more switches in the offspring than limiting familial information to steps I and IIa and then applying step III. This is due to the fact that the diploid FPM/LHCM applied to descendent individuals extracts linkage information to reconstruct the descendent padumnal haplotypes, yet performs better as it simultaneously accounts for the madumnal haplotype structure using population information.

We then examined the performances of DualPHASE and DAGPHASE when applied to real data. With these data, the true phase is unknown. To evaluate the performances of both programs in capturing LD information, we first phased the real data using familial information only. We selected 1183 sons with more than 10 half-sibs and reconstructed their paternal and maternal haplotypes using type I, IIA, and IIb information. As shown in Table 1, this family size allows phasing of more than 90% of the sons’ heterozygous markers with a switch rate less than 1/3000. We then evaluated how well DualPHASE and DAGPHASE would succeed in reconstructing these haplotypes under the worst-case scenarios, *i.e.*, (a) without any familial information or (b) with type I Mendelian information only. To avoid overestimation of the performances we used cross-validation. The 1183 sons with ≥ 10 half-sibs were split in 20 subsets. The FPM and LHCM models were trained on the complete data set (1701 bulls) minus one of the 20 subsets and used to reconstruct the haplotypes of the ignored subset. This was repeated 20 times. Table 2 summarizes the results. In the complete absence of familial information, switch rates were of the order of 1/65 (LHCM) and 1/40 (FPM) and of 1/135 (LHCM) and 1/75 (FPM) for the sparse and dense map segments, respectively. Extracting Mendelian information (I) decreased the overall switch rate to 1/145 (LHCM) and 1/120 (FPM) and to 1/240 (LHCM) and 1/240 (FPM) for the sparse

TABLE 1
Phasing accuracy—simulated data

Type	Familial information (steps I and II)		+ Population information (step III)			Overall	
	Phased Het SNPs	Switches	Phased Het SNPs	Switches FPM	Switches LHCM	Switches FPM	Switches LHCM
None (sires, <6 sons)	0 (0.0%) ^a	0.0 (0.0%)	687 (100.0%)	8.2 (1.2%)	3.7 (0.5%)	8.2 (1.2%)	3.7 (0.5%)
	0 (0.0%) ^b	0.0 (0.0%)	3525 (100.0%)	24.9 (0.7%)	4.0 (0.1%)	24.9 (0.7%)	4.0 (0.1%)
I (sons, <5 sibs)	334 (49.3%)	0.0 (0.0%)	343 (50.7%)	2.3 (0.7%)	1.8 (0.5%)	2.3 (0.3%)	1.8 (0.3%)
	1725 (49.5%)	0.0 (0.0%)	1763 (50.5%)	6.1 (0.3%)	3.6 (0.2%)	6.1 (0.2%)	3.6 (0.1%)
IIa (sires, ≥6 sons)	673 (98.3%)	0.1 (0.0%)	12 (1.7%)	0.0 (0.0%)	0.0 (0.0%)	0.1 (0.0%)	0.1 (0.0%)
	3462 (98.3%)	0.1 (0.0%)	60 (1.7%)	0.1 (0.2%)	0.0 (0.0%)	0.2 (0.0%)	0.2 (0.0%)
IIa (sires, >10 sons)	684 (99.9%)	0.0 (0.0%)	0.3 (0.1%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)
	3525 (99.9%)	0.0 (0.0%)	1.4 (0.1%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)
I+IIa (sons, ≥5 sibs)	334 (49.3%)	0.0 (0.0%)	343 (50.7%)	0.1 (0.0%)	0.1 (0.0%)	0.1 (0.0%)	0.1 (0.0%)
	1773 (50.1%)	0.0 (0.0%)	1763 (49.9%)	0.3 (0.0%)	0.2 (0.0%)	0.3 (0.0%)	0.2 (0.0%)
I+IIa (sons, ≥10 sibs)	334 (49.3%)	0.0 (0.0%)	343 (50.7%)	0.1 (0.0%)	0.1 (0.0%)	0.1 (0.0%)	0.1 (0.0%)
	1724 (49.4%)	0.0 (0.0%)	1764 (50.6%)	0.2 (0.0%)	0.2 (0.0%)	0.2 (0.0%)	0.2 (0.0%)
I+IIa+IIb (sons, ≥5 sibs)	617 (91.1%)	3.0 (0.5%)	60 (8.9%)	0.1 (0.2%)	0.1 (0.2%)	3.1 (0.5%)	3.1 (0.5%)
	3237 (92.9%)	13.1 (0.4%)	249 (7.1%)	0.2 (0.1%)	0.1 (0.0%)	13.2 (0.4%)	13.1 (0.4%)
I+IIa+ IIb (sons, ≥10 sibs)	617 (91.1%)	0.2 (0.0%)	60 (8.9%)	0.1 (0.2%)	0.0 (0.0%)	0.3 (0.0%)	0.2 (0.0%)
	3240 (92.9%)	0.5 (0.0%)	248 (7.1%)	0.1 (0.0%)	0.0 (0.0%)	0.7 (0.0%)	0.6 (0.0%)

Information used for phasing: I, use of Mendelian segregation rules to reconstruct the offspring's phase; IIa, use of linkage information to reconstruct the parental phase (marker alleles that tend to be cotransmitted reside on the same homolog); IIb, use of linkage information to reconstruct the offspring's phase (a chromosome segment flanked by two alleles originating from the same parental homolog is assumed to completely derive from that homolog unless the probability of double cross-over is > 0.999); III, use of population-wide linkage disequilibrium information.

^aLow-density map (2364 markers).

^bHigh-density map (12,113 markers). Of the markers corresponding to 1685 and 8623 markers for the low- and high-density maps, respectively, 71% were homozygous and considered phased de facto. The numbers shown pertain to remaining heterozygous markers only. In all instances, the FPM and LHCM models were trained on a haploid set of base haplotypes obtained by extracting type I information for all sons and type IIa information for sires with 6 or more sons.

and dense map segments, respectively. Thus, switch rates observed with the real dense map (80 markers/Mb) were similar to those observed with the simulated sparse map (23 markers/Mb). These results point toward some decrease in haplotyping reliability when working with real data. The number of switches, however, remained very acceptable, especially given the fact that in reality familial information will be available and utilized for most individuals. As for the simulated data, the LHCM model outperformed the FPM model to some extent.

Genotype imputation: An important feature of the proposed model is that it allows imputation of missing genotypes (*e.g.*, MARCHINI *et al.* 2007). Indeed, once the most likely HS of a given haplotype has been determined, missing marker information can be projected from the HS on the haplotype using the corresponding emission probabilities. To evaluate our imputation accuracy, we analyzed the real data set as follows. Haplotypes of the 1183 sons with >10 sibs were reconstructed on the sole basis of familial information (see above). Of the markers 1, 10, or 50% were randomly selected and the corresponding genotypes erased from the maternal haplotypes for a subset corresponding to $1183/20 \approx 60$ sons. The FPM and LHCM models were

trained on the base haplotypes of the rest of the animals (*i.e.*, total data set minus the subset of sons that were singled out) and subsequently used to impute the missing genotypes of the ~ 60 trimmed maternal haplotypes (imputation was done on the basis of LD only). Imputed genotypes were then compared with the actual ones and errors counted. The process was repeated for 19 such subsets until all sons would have been evaluated as 1 of 20 test sets and overall error rates compiled. Table 3 summarizes the results for the BTA14 segments with sparse and dense marker coverage. For the sparse map, with marker density slightly inferior to presently available $\sim 50K$ genome-wide SNP panels (CHARLIER *et al.* 2008; VAN TASSELL *et al.* 2008), imputation errors were $\sim 2\%$ when 1% of genotypes were missing, *i.e.*, comparable to the genotyping failure rate under routine conditions. Remarkably, imputation errors remained $<5\%$ even when imputing half the genotypes purely from LD information. For the dense part of the map, with coverage corresponding to a genome-wide panel of 250,000 SNP, imputation error remained less than 1% even with 50% missing genotypes. As for haplotyping, LHCM slightly outperformed FPM.

Haplotype clustering: Using FPM and LHCM, and assuming a specific map position, every haplotype in the

TABLE 2
Phasing accuracy—real data

Type	Familial information (steps I)		+ Population information (step III)			Overall	
	Phased Het SNPs	Switches	Phased Het SNPs	Switches FPM	Switches LHCM	Switches FPM	Switches LHCM
None	0 (0.0%) ^a	0.0 (0.0%)	65.22 (100.0%)	1.71 (2.6%)	0.98 (1.5%)	1.71 (2.6%)	0.98 (1.5%)
	0 (0.0%) ^b	0.0 (0.0%)	116.87 (100.0%)	1.52 (1.3%)	0.86 (0.7%)	1.52 (1.3%)	0.86 (0.7%)
I	35.80 (54.9%)	0.0 (0.0%)	29.42 (45.1%)	0.55 (1.9%)	0.45 (1.5%)	0.55 (0.8%)	0.45 (0.7%)
	64.21 (54.9%)	0.0 (0.0%)	52.66 (45.1%)	0.49 (0.9%)	0.48 (0.9%)	0.49 (0.4%)	0.48 (0.4%)

Information used for phasing: I, use of Mendelian segregation rules to reconstruct the offspring's phase; III, use of population-wide linkage disequilibrium information.

^aLow-density segment (256 markers).

^bHigh-density map (500 markers). Approximately 67 and 69% of the markers, corresponding to 171.08 and 343.46 markers for the low- and high-density maps, respectively, were homozygous and considered phased de facto. Approximately 8% of the markers, corresponding to 19.71 and 39.66 markers for the low- and high-density maps, respectively, remained unphased after step IIB and were not used for comparison.

population can be assigned to the “most likely” out of K hidden states. Hence, haplotypes are sorted in clusters of what are assumed to be genealogically related haplotypes. We verified this assertion on both simulated and real data.

We first used the *ms* software (HUDSON 2002) to simulate 1000 10-Mb haplotypes originating from a population with effective size of 100. As before, mutation rate was set at 10^{-8} per base and per generation, while the recombination rate was set at 1 cM/Mb. All polymorphisms with MAF >5% were retained for analysis, corresponding to 119 polymorphisms on average (or 12 SNPs/Mb). For each pair of haplotypes, we retrieved the time of coalescence at position 5 Mb, *i.e.*, in the middle of the chromosome (using the $-T$ option of *ms* to obtain gene trees, Newick format, and branch lengths). We consider that the time of coalescence is the best measure of relatedness between two chromosomes. We then applied the FPM and LHCM on the collection of 1000 haplotypes and determined for each of them the most likely HS at position 5 Mb. Two haplotypes were grouped in the same cluster at position 5 Mb only if they were assigned to identical HS at the two flanking markers. For comparison we computed all pairwise probabilities of IBD conditional on flanking marker data at the same position using the widely used deterministic approach of MEUWISSEN and GODDARD (2001), followed by the clustering of chromosomes with pairwise IBD probabilities >0.9 following DRUET *et al.* (2008), jointly referred to as “IBD approach.” To compute IBD probabilities, the effective population size was set at 100, the number of generations to the base population at 100, and information was considered for 5 markers on the left and right of the analyzed map position.

Table 4 summarizes the results. The probability for two randomly selected chromosomes to be part of the same cluster was higher for the IBD approach (0.2) than for FPM (0.12) and LHCM (0.08), suggesting that a larger proportion of chromosomes is included in a

smaller number of clusters with the IBD approach. The average IBD probability between chromosomes of the same cluster was higher for the IBD approach (0.97) than for the FPM (0.91) and LHCM (0.93) approaches, suggesting superior performances of the former. However, when using time of coalescence as measure of relatedness, the average time of coalescence (expressed in $4N$ generations) for pairs of clustered chromosomes proved to be lowest for LHCM (0.10 = 40 generations), intermediate for FPM (0.14 = 56 generations), and highest for IBD (0.21 = 84 generations). The figures in Table 4 suggest that the IBD approach tends to underestimate the within-LHCM/FPM cluster IBD probabilities. This may be due to the reduced ability of the IBD approach to deal with mutational events, which were not modeled by MEUWISSEN and GODDARD (2001). Thus, a mutation (or a genotyping error) in one of the two confronted haplotypes will heavily penalize the IBD probability irrespective of their resemblance on either side of the mutated site, while the LHCM/FPM models will be much more forgiving in such situation. Alternatively, the IBD approach may overestimate the within-IBD cluster IBD probabilities: the coalescence of haplotypes that are IBS over long segments may nevertheless be remote.

TABLE 3
Imputation accuracy—real data

Map segment	No. of markers	Missing markers	Error rate	
			FPM	LHCM
Sparse	256 12 SNPs/Mb	3 (1.2%)	0.023	0.021
		25 (9.8%)	0.025	0.022
		125 (48.8%)	0.046	0.031
Dense	500 84 SNPs/Mb	5 (1.0%)	0.006	0.007
		50 (10.0%)	0.007	0.007
		250 (50.0%)	0.015	0.010

TABLE 4

Genetic relatedness of clustered haplotypes measured by coalescence time or probability of identity-by-descent: simulations

Clustering algorithm	% clustered chrom. pairs	Time of coalescence ($4N$ generations)		Probability of identity-by-descent (IBD)	
		Within clusters	Between clusters	Within clusters	Between clusters
LHCM	8.2	0.10	1.11	0.93	0.27
FPM	12.1	0.14	1.15	0.91	0.24
IBD	19.9	0.21	1.23	0.97	0.16

We then compared the clustering of the 1183 real BTA14 haplotypes obtained with the IBD, LHCM, or FPM approach. Dense and sparse map segments were considered separately. As for the simulated data, the IBD approach clustered more pairs of haplotypes (20 and 12% for the dense and sparse maps, respectively) than FPM (15 and 9%) and LHCM (11 and 7%) and the mean IBD probability between grouped haplotypes was higher (0.98 and 0.98) than with FPM (0.92 and 0.84) and LHCM (0.91 and 0.76). Figure 1 shows the cumulative distributions (in percentage) of grouped (or non-grouped) haplotype pairs with increasing IBD probability for the three methods. For the dense map (Figure 1A), 87.4% (FPM), 86.5% (LHCM), and 100.0% (IBD approach) of grouped haplotypes have IBD probability higher than 0.90. With all methods, more than 50% of grouped haplotypes have IBD probability exceeding 0.99. However, the figure indicates that a nonnegligible proportion of haplotype pairs grouped by FPM or LHCM have low IBD probability. This is due to the higher tolerance of FPM/LHCM to punctual differences (genotyping errors and mutations) and to “edge effects”; *i.e.*, transitions between hidden states occur abruptly at specific marker positions with FPM/LHCM while IBD probabilities evolve gradually. Figure 1A also shows that for a majority ($\sim 70\%$) of nongrouped haplotypes, the estimated IBD probability is zero irrespective of the method. However, although the IBD approach by definition groups all pairs with IBD probability exceeding 0.90, with FPM and LHCM approximately 10% of nongrouped haplotype pairs have IBD probability >0.90 . Indeed, with FPM/LHCM haplotypes that are identical-by-state over ~ 10 markers may sometimes be assigned to distinct clusters differing further down the map. Finally, Figure 1B shows that for the lower density map, more haplotype pairs have intermediate IBD probabilities resulting in more grouped haplotypes with IBD probability lower than 0.90 and fewer nongrouped pairs with IBD probability of either 0 or >0.90 .

QTL mapping: Having demonstrated that the FPM and LHCM cluster genealogically related haplotypes as expected, we evaluated the utility of this information to map and fine map QTL. To that end, we analyzed the real pedigree, phenotype, and BTA14 genotype data, knowing that BTA14 harbors the *DGATI* gene at map

position 445 kb and that the examined population segregates for the previously described *Lys232Ala* mutation that has a major effect on milk fat composition (GRISART *et al.* 2002, 2004). The frequency of the fat increasing allele ($=Lys$) in the maternal chromosomes of the sons was 0.46. QTL mapping was performed after exclusion of *DGATI* polymorphisms from the SNP data. The closest SNPs were respectively at map position 291 Kb (SNP 20) and 560 Kb (SNP 21). Haplotype clusters defined by either of the three approaches (FPM, LHCM, IBD) were included as additive random effects in a mixed model as detailed in MATERIALS AND METHODS.

Overall the three methods yielded very similar location score profiles (Figure 2). The curves obtained with FPM and LHCM were, in general, smoother than with the IBD approach. As expected the three curves maximized in the vicinity of the *DGATI* gene. Note that the IBD approach computes location scores in marker intervals, while the FPM and LHCM approaches compute location scores at marker positions. The IBD curve correctly maximized in the interval containing *DGATI*. Likewise, with the FPM approach the two most significant markers (SNP 20 and 21) were adjacent and immediately flanking *DGATI*. With the LHCM approach, however, the two most significant markers were 22 and 23 when analyzing the chromosome from left to right and were 19 and 20 when analyzing from right to left. The corresponding intervals are respectively 135 and 139 kb away from *DGATI*. We attribute this lesser precision of LHCM to the inherent difficulties in reliably reconstructing the extremities of the DAG. The maximum LRT was highest for the IBD approach, followed by the LHCM and FPM approaches. However, differences were modest.

QTL mapping power and accuracy should be determined by the ability to segregate the causative QTN variants in distinct haplotype clusters, *i.e.*, the level of LD between haplotype clusters and QTN as measured by r^2 . Knowledge of the causative *Lys232Ala DGATI* mutation in the case of the BTA14 QTL on milk fat content allowed us to compare the discriminating performances of the three approaches. Following HAYES *et al.* (2006), the proportion of the QTN variance explained by the haplotype clusters was computed as

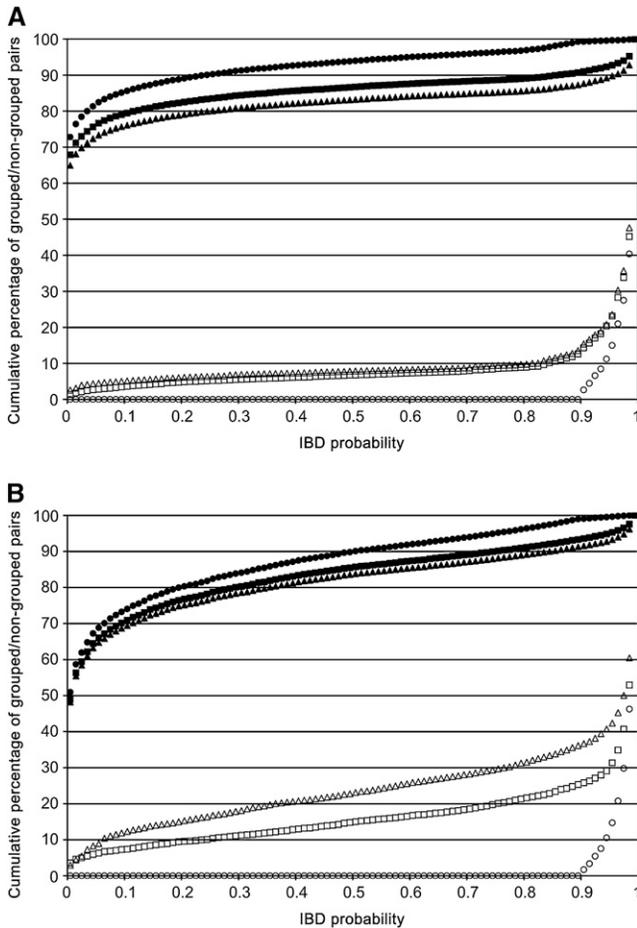


FIGURE 1.—Percentage (y-axis) of grouped (open symbols) or nongrouped (solid symbols) haplotype pairs with IBD probability below a certain probability (x-axis). FPM (squares), LHCM (triangles), or IBD approach (circles). It can, for instance, be seen that 100% of haplotype pairs that are not grouped with the IBD approach have IBD probability < 0.9 (solid circle) and that 0% of haplotype pairs that are grouped with the IBD approach have IBD probability < 0.9 (open circle). (A) dense part of BTA14, (B) sparse part of BTA14.

$$r^2 = \frac{1}{q(1-q)} \sum_{i=1}^K \frac{D_i^2}{p_i},$$

where q is the population frequency of the QTN (assuming a biallelic QTL), p_i is the population frequency of haplotype cluster i (of K), and D_i is the LD between haplotype cluster i , and the QTN computed as $p_i(q | p_i) - p_i q$. Values of r^2 , computed at the maximum LRT position for each method, were 0.62, 0.62, and 0.68 for FPM, LHCM, and IBD, respectively, hence suggesting a slight superiority of the IBD approach in agreement with the location scores. This higher r^2 was, however, obtained with twice as many clusters or more.

Figure 3 shows for the 20, 16, and 40 haplotype clusters defined respectively by the FPM, LHCM, and IBD approaches at their respective maximum LRT location (i) population frequency, (ii) composition with

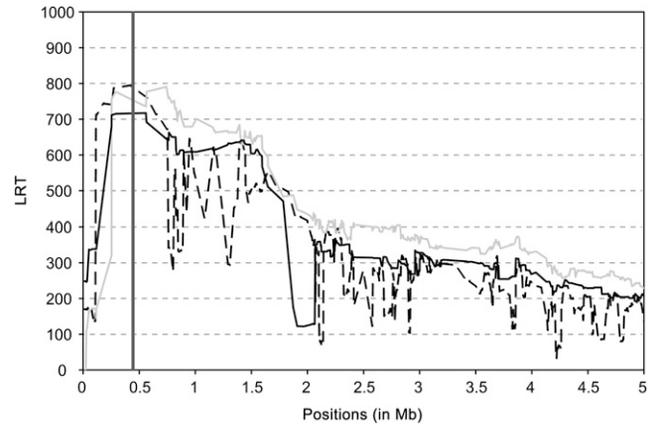


FIGURE 2.—LRT profiles in the first 5 Mb of BTA14 for fat content obtained with L + LD fine-mapping methods: FPM (solid line), LHCM (shaded line), and IBD approach (dashed line). DGAT1 position is indicated by the vertical line.

respect to the *DGAT1* *Lys232Ala* mutation, and (iii) effect on fat content. The three methods yielded congruent haplotype patterns confirming (i) the major effect of the *Lys232Ala* mutation and (ii) the previously described paradoxical haplotype homogeneity of the ancestral *Lys232* allele [$\sim 79\%$ (FPM), 78% (LHCM) and 76% (IBD) of the *Lys232* alleles are confined to one haplotype cluster] *vs.* haplotype heterogeneity of the derived *232Ala* allele [7 (HMM), 7 (LHCM), and 9 (IBD) haplotype clusters are required to account for 80% of the *232Ala* alleles] (GRISART *et al.* 2002).

DISCUSSION

We herein describe a method and accompanying software that efficiently phases high-density SNP genotype data. The proposed method builds on either of two previously described hidden Markov models (FastPHASE/FPM, SCHEET and STEPHENS 2006; Beagle/LHCM, S. R. BROWNING and B. L. BROWNING 2007) implemented to phase genotypes from human populations. An important advance of our approach is that it simultaneously extracts population (LD) and familial information (Mendelian segregation and linkage), whereas previous methods used either one of these but not both (note that the latest Beagle version handles trios; BROWNING and BROWNING 2009). This feature improves the accuracy of the phase reconstruction (see Tables 1 and 2). It particularly affects the integrity of long-range haplotypes. This knowledge may be useful for the detection of founder mutations on the basis of shared extended haplotypes.

Being based on hidden Markov models, the developed methods assign reconstructed haplotypes to a limited number of hidden clusters. We propose direct use of the ensuing information for QTL mapping in a mixed model/REML setting. We demonstrate that both

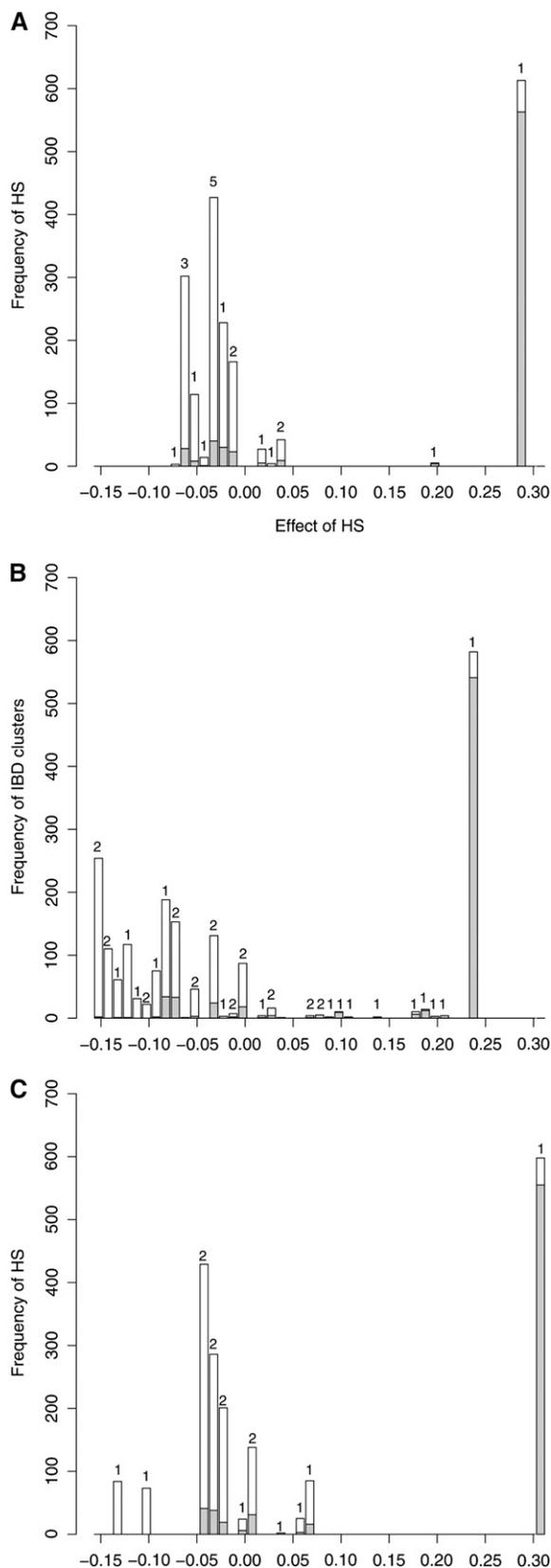


FIGURE 3.—Frequency and composition with respect to the *DGATI Lys232Ala* mutation (shaded and open for the ancestral and derived alleles, respectively) and effect on fat content (in percentage of fat) of the haplotype clusters defined by FPM (A), LHCM (B), and the IBD approach (C). The number above each bar refer to the corresponding numbers of clusters.

FPM and LHCM clusters generate mapping results that are comparable to those obtained with the standard approach based on pairwise IBD probabilities (MEUWISSEN and GODDARD 2001), yet with substantial computational gain. On a dual-core computer (running at 2.13 GHz and with 2 GB RAM), DAGPHASE (LHCM) would phase and cluster the haplotypes corresponding to the real BTA14 data in 47 min and DualPHASE (FPM) in 966 min, while the IBD approach implemented with our own program would require 9133 min to compute all pairwise IBD probabilities using already phased genotypes. The corresponding output files were 28 Mb for DAGPHASE, 21 Mb for DualPHASE, and 74 Gb for the IBD approach. We used IBD probabilities as in MEUWISSEN and GODDARD (2001) although MEUWISSEN and GODDARD (2007) added mutation to their model. In addition we limited the computation of IBD probabilities to windows of 10 markers although the method could be applied to longer markers windows, at the expense of increased computing time.

In this study, we applied only single QTL models. However, it is straightforward to extend the approach to multiple QTL, including fitting as many effects as available SNPs as done in genomic selection (MEUWISSEN *et al.* 2001). Analyses to evaluate the benefit of haplotype rather than single marker information in genomic selection are ongoing (A. P. W. DE ROOS and T. DRUET, unpublished results).

One advantage of clustering haplotypes on the basis of their relatedness followed by the estimation of the phenotypic effect of individual clusters is the ensuing possibility of identifying clusters that are most likely to be functionally different and on which to focus molecular studies to identify the causative QTL.

The proposed approach allows accurate genotype imputation (see Table 3). For missing data points, the most likely genotypes can be inferred from the emission probabilities of the clusters to which an individual's haplotypes are assigned. The ability to reliably fill in missing genotypes facilitates the computation of pairwise IBD probabilities using standard methods (*e.g.*, MEUWISSEN *et al.* 2001; DRUET *et al.* 2008). It also allows merging of data sets generated with distinct SNP panels (*e.g.*, MARCHINI *et al.* 2007; T. DRUET, unpublished results). This application is relevant when part of the population is genotyped using high-density arrays yet the majority is genotyped with a low-density panel. Such schemes have been proposed for the fine mapping of QTL using nested association mapping in experimental crosses of plants (YU *et al.* 2008), as well as for expanding genomic selection to the general dairy cattle population (HABIER *et al.* 2009). Moreover, the ability to predict the most likely genotype from hidden state information can easily be exploited for the detection of likely genotyping errors.

The LinkPHASE (steps I and II), HiddenPHASE (step III with FPM), DualPHASE (steps I, II, and III with FPM)

and DAGPHASE (step III with LHCM) programs form the PHASEBOOK programs suite and are available from the authors upon request.

The authors thank Paul Scheet and Sharon Browning for answering our questions on fastPHASE and Beagle and Wouter Coppieters, Frédéric Farnir and Mathieu Gautier, for their comments and suggestions on this work. Tom Druet is Research Associate from the Fonds National de la Recherche Scientifique (FNRS). This work was funded by grants from the Service Public de Wallonie and from the Communauté Française de Belgique (Biomod ARC). CRV (<http://www.crv4all.com/>) provided the real data set.

LITERATURE CITED

- BLOTT, S., J. J. KIM, S. MOISIO, A. SCHMIDT-KUNTZEL, A. CORNET *et al.* 2003 Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. *Genetics* **163**: 253–266.
- BROWNING, B. L., and S. R. BROWNING, 2007 Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet. Epidemiol.* **31**: 365–375.
- BROWNING, B. L., and S. R. BROWNING, 2009 A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**: 210–223.
- BROWNING, S. R., 2006 Multilocus association mapping using variable-length Markov chains. *Am. J. Hum. Genet.* **78**: 903–913.
- BROWNING, S. R., and B. L. BROWNING, 2007 Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**: 1084–1097.
- CHARLIER, C., W. COPPIETERS, F. ROLLIN, D. DESMECHT, J. S. AGERHOLM *et al.* 2008 Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nat. Genet.* **40**: 449–454.
- DRUET, T., S. FRITZ, M. BOUSSAHA, S. BEN-JEMAA, F. GUILLAUME *et al.* 2008 Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map. *Genetics* **178**: 2227–2235.
- GEORGE, A. W., P. M. VISSCHER and C. S. HALEY, 2000 Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics* **156**: 2081–2092.
- GODDARD, M.E., and B.J. HAYES, 2009 Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.* **10**(6): 381–391.
- GRISART, B., W. COPPIETERS, F. FARNIR, L. KARIM, C. FORD *et al.* 2002 Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Res.* **12**: 222–231.
- GRISART, B., F. FARNIR, L. KARIM, N. CAMBISANO, J. J. KIM *et al.* 2004 Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc. Natl. Acad. Sci. USA* **101**: 2398–2403.
- GUO, S. W., 1994 Computation of identity-by-descent proportions shared by two siblings. *Am. J. Hum. Genet.* **54**: 1104–1109.
- HABIER, D., R. L. FERNANDO and J. C. DEKKERS, 2009 Genomic selection using low-density marker panels. *Genetics* **182**: 343–353.
- HAYES, B. J., A. J. CHAMBERLAIN and M. E. GODDARD, 2006 Use of markers in linkage disequilibrium with QTL in breeding programs. Proceeding from the 8th World Congress of Genetics Applied to Livestock Production, Belo Horizonte, Brazil, August 13–18.
- HENDERSON, C. R., 1984 Estimation of variances and covariances under multiple traits models. *J. Dairy Sci.* **67**: 1581–1589.
- HUDSON, R. R., 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**: 337–338.
- JOHNSON, D. L., and R. THOMPSON, 1995 Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *J. Dairy Sci.* **78**: 449–456.
- KONG, A., G. MASSON, M. L. FRIGGE, A. GYLFASON, P. ZUSMANOVICH *et al.* 2008 Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**: 1068–1075.
- MARCHINI, J., B. HOWIE, S. MYERS, G. McVEAN and P. DONNELLY, 2007 A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**: 906–913.
- MEUWISSEN, T. H., and M. E. GODDARD, 2000 Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* **155**: 421–430.
- MEUWISSEN, T. H., and M. E. GODDARD, 2001 Prediction of identity by descent probabilities from marker-haplotypes. *Genet. Sel. Evol.* **33**: 605–634.
- MEUWISSEN, T. H., and M. E. GODDARD, 2007 Multipoint identity-by-descent prediction using dense markers to map quantitative trait loci and estimate effective population size. *Genetics* **176**: 2551–2560.
- MEUWISSEN, T. H., B. J. HAYES and M. E. GODDARD, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–1829.
- RABINER, L. R., 1989 A tutorial on Hidden Markov Chains and selected applications in speech recognition. *Proc. IEEE* **77**: 257–286.
- SCHEET, P., and M. STEPHENS, 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**: 629–644.
- VANRADEN, P. M., and G. R. WIGGANS, 1991 Derivation, calculation, and use of national animal model information. *J. Dairy Sci.* **74**: 2737–2746.
- VAN TASSELL, C. P., T. P. SMITH, L. K. MATUKUMALLI, J. F. TAYLOR, R. D. SCHNABEL *et al.* 2008 SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* **5**: 247–252.
- VOIGHT, B.G., S. KUDARAVALLI, X. WEN and J.K. PRITCHARD, 2006 A map of recent positive selection in the human genome. *PloS Biol.* **4**(3): e72.
- Wellcome Trust Case Control Consortium, 2007 Genome wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**: 645–646.
- WINDIG, J. J., and T. H. E. MEUWISSEN, 2004 Rapid haplotype reconstruction in pedigrees with dense marker maps. *J. Anim. Breed. Genet.* **121**: 26–39.
- YU, J., J. B. HOLLAND, M. D. McMULLEN and E. S. BUCKLER, 2008 Genetic design and statistical power of nested association mapping in maize. *Genetics* **178**: 539–551.

Communicating editor: H. ZHAO

GENETICS

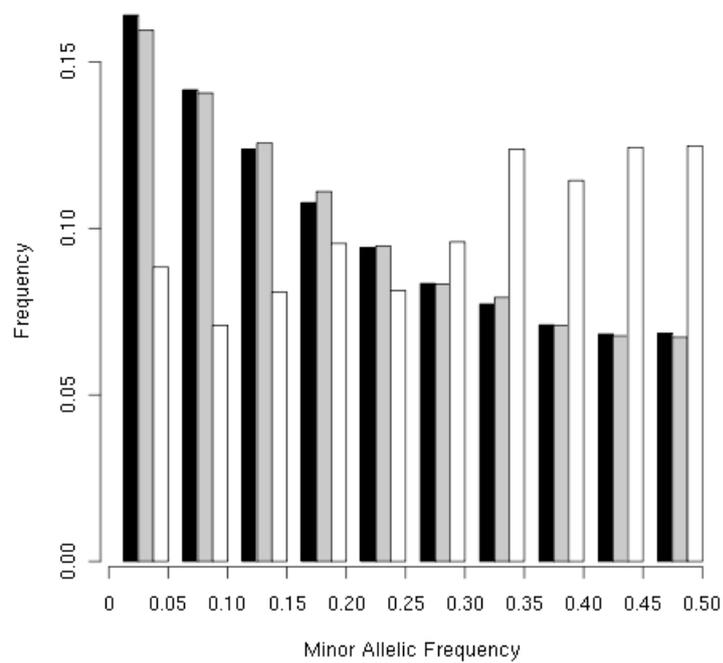
Supporting Information

<http://www.genetics.org/cgi/content/full/genetics.109.108431/DC1>

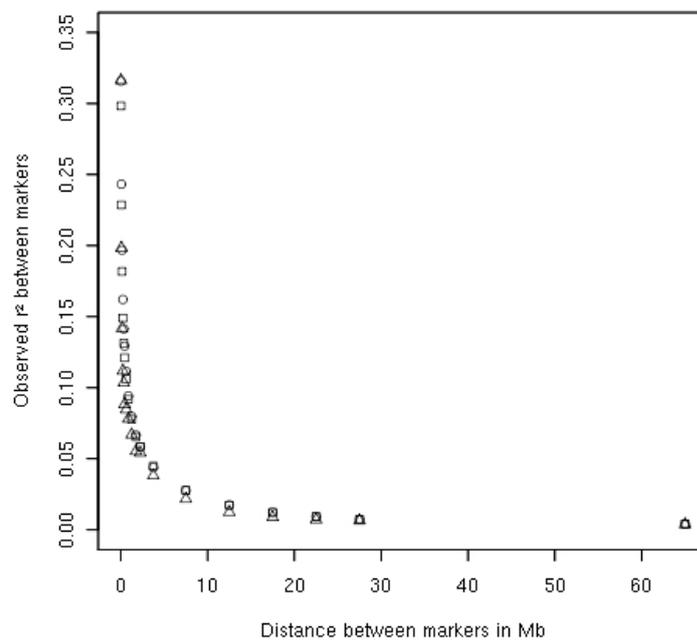
**A Hidden Markov Model Combining Linkage and Linkage
Disequilibrium Information for Haplotype Reconstruction
and Quantitative Trait Locus Fine Mapping**

Tom Druet and Michel Georges

Copyright © 2010 by the Genetics Society of America
DOI: 10.1534/genetics.109.108431



A



B

FIGURE S1.—(A) Frequency (in percent) of classes of minor allelic frequencies, (B) mean r^2 between markers (results average by distance between markers). Results were compiled for simulated data set (sparse map in black or circles; dense map in grey or squares) and real data set (in white or triangles).