BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 19 2012, pages 2467–2473
doi:10.1093/bioinformatics/bts348

Genetics and population analysis

Advance Access publication June 17, 2012

Ancestral haplotype-based association mapping with generalized linear mixed models accounting for stratification

Z. Zhang[1], F. Guillaume[1], A. Sartelet[1], C. Charlier[1], M. Georges[1], F. Farnir[2,†] and T. Druet[1,*,†]

[1]Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, University of Liège (B34), 1 avenue de l'Hôpital and [2]Unit of Animal Productions, Faculty of Veterinary Medicine, University of Liège (B43), 20 boulevard de Colonster, B-4000 Liège, Belgium

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: In many situations, genome-wide association studies are performed in populations presenting stratification. Mixed models including a kinship matrix accounting for genetic relatedness among individuals have been shown to correct for population and/or family structure. Here we extend this methodology to generalized linear mixed models which properly model data under various distributions. In addition we perform association with ancestral haplotypes inferred using a hidden Markov model.

Results: The method was shown to properly account for stratification under various simulated scenari presenting population and/or family structure. Use of ancestral haplotypes resulted in higher power than SNPs on simulated datasets. Application to real data demonstrates the usefulness of the developed model. Full analysis of a dataset with 4600 individuals and 500 000 SNPs was performed in 2 h 36 min and required 2.28 Gb of RAM.

Availability: The software GLASCOW can be freely downloaded from www.giga.ulg.ac.be/jcms/prod_381171/software.

Contact: francois.guillaume@jouy.inra.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 24, 2012; revised on May 24, 2012; accepted on June 12, 2012

1 INTRODUCTION

Genome-wide association studies (GWASs) identify genetic variants (e.g. SNPs, CNV or indels) affecting traits of interest such as those related to human health or of agronomical importance. With the development of high-throughput genotyping and next-generation sequencing, these studies have been particularly successful. Hundreds of loci associated with diseases were detected through GWAS (e.g. Donnelly, 2008). Association studies proved equally valuable in other organisms such as Arabidopsis thaliana (Aranzana *et al*., 2005), mice (Threadgill *et al*., 2002), dog, crops (Malosetti *et al*., 2007; Yu *et al*., 2006) or livestock species.

Although very effective, genetic association studies still face a number of potential pitfalls. One major problem in GWAS comes from the spurious associations that may occur as a result of relatedness between individuals (e.g. familial relationships or population structure). Another issue is that, especially for complex traits, non-genetic factors (e.g. sex, age, etc.) may have profound impact on the scrutinized phenotype, raising the need for proper modeling of these effects.

An appealing solution to these problems is to use a mixed-model framework. Indeed, this methodology makes it possible to include covariates in the model and to account for the average genomic relatedness among individuals (population or family structure). Such models have been used for many years for QTL mapping especially in animal breeding (George *et al*., 2000). Recent studies (Kang *et al*., 2008; Malosetti *et al*., 2007; Yu *et al*., 2006; Zhao *et al*., 2007) have demonstrated that inclusion of such effects in mixed-models properly corrects for stratification and that the use of mixed models to control for stratification resulted in fewer false positives and/or higher power than other techniques such as genomic control (Devlin and Roeder, 1999), structured association (Pritchard *et al*., 2000) or principal components analysis (Price *et al*., 2006). In addition, mixed-models were able to capture the multiple levels of population structure and genetic relatedness. All these features make mixed-models a very promising tool to perform association analyses while controlling for relatedness structure.

Linear mixed-models (LMMs) assume that traits are normally distributed. Use of generalized linear mixed models (GLMMs) allows extension of the mixed-model approach to other types of traits, such as binary traits for example. With these models, a linear function of different covariates including polygenic and local genomic effects is used to describe the expected value of the observed phenotype through a so-called link function. Tzeng and Zhang (2007) developed a variance-components score test for association studies which can be used in the GLMM framework.

Analyses can be performed using single SNP or haplotypes of multiple SNPs. Haplotypes are specific combinations of alleles on the same chromosome. They extract more information on the relation between DNA variation and phenotypes than single SNPs and may present higher correlation with underlying mutations [depending on the marker density and the linkage disequilibrium (LD) pattern in the population]. Furthermore, haplotype tests can model allelic heterogeneity or find several (interacting) mutations at different tightly linked sites. However, the power of haplotype-based tests is potentially reduced due to the extra

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.

degrees of freedom needed in these analyses (e.g. Su *et al*., 2008; Tzeng and Zhang, 2007). Different strategies to minimize this problem have been proposed in the literature, relying mainly on grouping haplotypes based on similarity (e.g. Blott *et al*., 2003; Durrant *et al*., 2004; Druet *et al*., 2008; Seltman *et al*., 2003). Various clustering algorithms are available. Those relying on sliding window approaches are not optimal as the optimal window size varies from one region to another (Browning, 2008). With the localized haplotype clustering method (Browning and Browning, 2007), clusters of haplotypes are parsimoniously selected. This model allows for greater flexibility because haplotype lengths and the number of haplotypes are variable. Browning (2008) stated that this model is conceptually similar to the clusters of the Scheet and Stephens (2006) model. For each position along the genome, this later model assigns haplotypes to a predetermined number of ancestral haplotypes present several generations ago from which all haplotypes within a cluster are assumed to have descended. Each haplotype can be associated to a cluster for a different length making the model flexible. In addition, the model can group haplotypes with small difference (missing genotypes, genotyping errors or recent mutations). Su *et al*. (2008) proposed to use these ancestral haplotypes in association testing, whereas we suggested to use them for QTL fine-mapping and genomic selection (Druet and Georges, 2010). In the same study, we showed that these clusters group haplotypes having a recent common ancestor (with short time to coalescence) and high identity-by-descent (IBD) probabilities [as estimated with the method of Meuwissen and Goddard (2001)]. It was as efficient as methods using these IBD probabilities to cluster haplotypes (e.g. Druet *et al*., 2008). The use of these ancestral haplotypes proved already efficient for QTL fine-mapping (Karim *et al*., 2011) and genomic selection (de Roos *et al*., 2011).

In the present study, we develop a haplotyped-based method for association mapping relying on GLMM accounting for stratification and other covariates affecting the modeled trait.

2 METHODS

The proposed method relies on GLMM. To account for stratification and/or polygenic background, the model includes a vector of random polygenic effects (e.g. Yu *et al*., 2006) in addition to the random haplotype effects:

$$\eta = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{H}\mathbf{h} \tag{1}$$

where η is a vector of n linear predictors (with n equal to the number of observed phenotypes), β is a vector of fixed effects including the overall mean, **u** is a vector of n' random polygenic effects (with n' equal to the number of individuals for which genomic information is available typically, $n' = n$), **h** is a vector of P random ancestral haplotype effects, [with P equal to the chosen number of ancestral haplotypes (Druet and Georges, 2010)]. **X**, **Z** and **H** are incidence matrices relating, respectively, fixed effects, polygenic and ancestral haplotype effects to the linear predictor. The variance of the random polygenic effects is denoted $\mathbf{G} = 2\,\mathbf{K}\,\sigma_g^2$ (Yu et al., 2006) where **K** is a relative kinship matrix obtained from the marker data (see below) and σ_g^2 is the additive genetic variance. The variance of the random ancestral haplotype effects, is denoted $\mathbf{V} = \mathbf{I}\,\sigma_h^2$ where σ_h^2 is the 'haplotypic' variance. The covariances between ancestral haplotype effects were assumed to be zero.

The linear predictors are transformed to the observed scale (e.g., disease status) through $h(\cdot)$, the inverse function of the link function $g(\cdot)$ (e.g. McCullagh and Nelder, 1989). In our analyses, the logit link function was used to model binomial data such as disease status. The probability for individual i to be affected by the disease $h(\eta_i)$ is therefore obtained through the inverse of the logit function:

$$\mu_i = h(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \tag{2}$$

Coding healthy individuals with 0 and affected ones with 1, this probability is also the expected value for the trait. The solutions of the GLMM were obtained using an iterative procedure based on the Laplace approximation of the likelihood (Breslow and Clayton, 1993; McCullagh and Nelder, 1989). Indeed, the GLMM equations can be approximated by the following mixed-model equations (MMEs):

$$\begin{bmatrix} \mathbf{X}'\mathbf{W}\mathbf{X} & \mathbf{X}'\mathbf{W}\mathbf{Z} & \mathbf{X}'\mathbf{W}\mathbf{H} \\ \mathbf{Z}'\mathbf{W}\mathbf{X} & \mathbf{Z}'\mathbf{W}\mathbf{Z} + \mathbf{G}^{-1} & \mathbf{Z}'\mathbf{W}\mathbf{H} \\ \mathbf{H}'\mathbf{W}\mathbf{X} & \mathbf{H}'\mathbf{W}\mathbf{Z} & \mathbf{H}'\mathbf{W}\mathbf{H} + \mathbf{V}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\beta} \\ \widehat{\mathbf{u}} \\ \widehat{\mathbf{h}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{W}\tilde{\mathbf{y}} \\ \mathbf{Z}'\mathbf{W}\tilde{\mathbf{y}} \\ \mathbf{H}'\mathbf{W}\tilde{\mathbf{y}} \end{bmatrix} \tag{3}$$

where **W** is a diagonal matrix with element $w_i = (\phi\upsilon(\mu_i)\{g'(\mu_i)\}^2)^{-1}$, with ϕ is a dispersion parameters and $\upsilon(\cdot)$ is a known variance function; $\widetilde{\mathbf{y}}$ is a vector with elements $\widetilde{y}_i = \eta_i + g'(\mu_i)(y_i - \mu_i)$. In the case of the logit function (and one record per individual), **W** is a diagonal matrix with $\mu_i(1 - \mu_i)$ on the diagonal. Since equations coefficients are functions of the unknowns, an iterative procedure must be used. First, starting values for β, **u** and **h** are used to compute μ and to build the approximate MME. These are then solved to obtain new estimates of β, **u** and **h**. Parameters of the model (including variance components) are repeatedly estimated by REML until convergence.

The relative kinship matrix **K** was estimated based on similarity scores as in Eding *et al*. (2001) and Hayes and Goddard (2008):

$$S_{XY,l} = 0.25[I_{11} + I_{12} + I_{21} + I_{22}] \tag{4}$$

where $S_{XY,l}$ is the similarity score between individuals x and y at locus l and I_{ij} is an indicator variable equal to 1 if allele i on locus l in the individual x and allele j on the same locus in individual y are identical, otherwise it is 0. In our analyses, we replaced SNP alleles by ancestral haplotypes (similar to multi-allelic markers). The relationships are then based on closer founders. S_{XY}, averaged over the whole genome and normalized is then used as an estimator of the kinship relationship f_{XY} as Eding et al. (2001):

$$\hat{f}_{XY} = \frac{S_{XY} - s}{1 - s} \tag{5}$$

where s is the minimum value of S_{XY} in the matrix (Hayes and Goddard, 2008). Kang et al. (2008) and Zhao et al. (2007) concluded that use of similarity score to construct relationship matrices was as efficient as more complex methods and avoided problems of non-positive definite matrices. We tested different methods to construct relationship matrices (based on SNP or haplotypes) but these had little impact on estimation of polygenic effects and even less on residuals.

Associations are tested for every marker position along the genome by a significance test of $\sigma_h^2 = 0$. Explicit evaluation of the likelihoods in GLMM is cumbersome, making application of likelihood ratio tests (LRTs) challenging. Therefore we used the score tests as proposed by Verbeke and Molenberghs (2003). Schaid *et al*. (2002) and Tzeng and Zhang (2007) used score tests in haplotype-based association studies with binary traits. The score tests are based on the value of the first derivative of the log-likelihood under the null hypothesis (i.e. the variance of the haplotypes is null). A significant positive first derivative with respect to the variance component indicates that the maximum-likelihood estimator of the haplotypes variance is significantly different from zero.

Tzeng and Zhang (2007) derived a test statistic T based on the score tests for haplotype-based models for GLMM. In the case of the logit function, the test statistic is equal to:

$$T = 0.5(\mathbf{y} - \mu)'\mathbf{H}\mathbf{H}'(\mathbf{y} - \mu) = 0.5(\mathbf{H}'(\mathbf{y} - \mu))' * \mathbf{H}'(\mathbf{y} - \mu) \tag{6}$$

where $\mathbf{y} - \mu$ is a vector of residuals (observations corrected for estimated fixed and random effects) obtained from a GLMM under the null hypothesis (no haplotype effect) where $\sigma_h^2 = 0$. Since T relies on estimation of residuals from a model without haplotype effects, the procedure is similar to the two-step procedure proposed in Aulchenko et al. (2007). Therefore it has the same advantages: the mixed models must be solved only once to obtain the residuals, which considerably speeds up computations, and since residuals are corrected for stratification, they are free from familial correlations and the data become exchangeable (Aulchenko et al., 2007) which means that permutation techniques may be applied.

Tzeng and Zhang (2007) demonstrated that the distribution of the T test statistic under the null hypothesis could be approximated using a gamma distribution. We perform 1,000 permutations of the residuals to estimate the mean and the variance of the gamma distribution (or the shape and scale parameters). Parameters of the gamma distribution are estimated for each tested position (marker) because the distribution is influenced by the structure of the incidence matrix relating haplotypes to residuals, which is potentially position specific. We will refer to this strategy as 'gamma approximation'. In addition, empirical *P*-values can be computed by repeatedly permuting the phenotypes (residuals) among the individuals (referred to as permutation hereafter).

3 SIMULATIONS

A dataset of 3547 genotyped Holstein, Jersey or crossbred bulls (Karim *et al.*, 2011) was used to simulate case/control studies. Individuals were genotyped for the Illumina Bovine SNP50 SNP chip (Illumina, San Diego, CA). After editing data, 37 647 SNPs were conserved on the 29 autosomes. DualPHASE from the PHASEBOOK package (Druet and Georges, 2010) was used to infer haplotypes from the genotyped individuals and to assign them to K ancestral haplotype clusters (K was set equal to 5, 10 or 20).

To simulate different stratification scenarios including breed (i.e. structured populations) and polygenic (i.e. familial relationships) effects, and mimicking major variant (SNP with large effect) effects, the following model was used:

$$\eta = \mathbf{X_1}\mu + \mathbf{X_2}\beta + \mathbf{Z_1}\mathbf{u} + \mathbf{Z_2}\mathbf{v} \tag{7}$$

where μ is the mean effect, $\mathbf{X_1}$ is a vector of '1', β is the breed effect, $\mathbf{X_2}$ is a vector containing the percentage of Holstein blood (ranging from 0 to 1), $\mathbf{Z_1}$ is a matrix ($n \times 1000$) containing the number of alleles '1' of a set of 1000 SNPs with small phenotypic effect (hereafter called polygenic SNPs), $\mathbf{u}$ is a vector containing the allelic substitution effects of 1000 polygenic SNPs used to simulate a polygenic effect, $\mathbf{Z_2}$ is a vector containing the number of alleles '1' for a SNP with a large phenotypic effect (hereafter called major variant) and $\mathbf{v}$ is the allelic substitution effect for that SNP. The breed effect was equal to 0, 0.2, 0.4 or 0.7 according to the scenario (corresponding to odds ratios (ORs) equal to 1.0, 1.22, 1.5 and 2.0, respectively), the individual SNP effects (1000 SNPs) were drawn from a gamma distribution (shape = 0.4 and scale = 1). The variance of the polygenic effects (the sum of 1000 individual polygenic SNP effects) was then rescaled to 0 or 0.16. Finally, the major variant effects were equal to 0.5 and 0.8 (corresponding to OR equal to 1.65 and 2.22, respectively). The phenotype of each individual was then sampled from a Bernoulli distribution with mean equal to:

$$p_i = h(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \tag{8}$$

The overall mean effect was set to obtain a prevalence of the disease of 27% in the population. Finally, 500 cases and 500 controls were randomly sampled from the 3547 genotyped individuals. One thousand SNPs were selected as potential major variants and were removed from the dataset prior to phasing. A total of 10 000 and 100 000 simulations were performed per scenario to estimate power and to compute QQ-plots, respectively.

4 RESULTS

4.1 Simulated data

The maximum LD (measured by r^2) between each of the 1000 major variants and the remaining SNPs (hereafter called marker SNPs) or with the 5, 10 or 20 ancestral haplotypes was estimated within a 2 Mb interval (1 Mb on each side of the major variant). The maximum r^2 was on average equal to 0.40 with marker SNPs and to 0.51 and 0.72 with 10 or 20 ancestral haplotypes. However, in some cases, SNPs can still present higher LD. Indeed, using marker SNPs or $K = 20$ haplotypes, 7.5 % of the major variants were captured with an r^2 of 1.0 whereas this value dropped to 0% with 5 or 10 haplotypes. From here on, $K = 20$ for the remainder of the simulation study.

To test whether our model correctly accounted for stratification, type-I errors were estimated by testing the model under H0 (the major variant had no effect: v was set equal to 0) and QQ-plots were generated. Simulations were performed including a breed effect of 0.4, but no polygenic effect. Four models were fitted to the generated dataset: using haplotypes or marker SNPs, and including or not a polygenic effect accounting for stratification. In Figure 1, models without polygenic effect clearly present an excess of small *P*-values. After inclusion of the polygenic effect, the regression slopes of the QQ-plot were below one, indicating that stratification was correctly accounted for but that tests are slightly too conservative.

In Table 1, regression coefficients of QQ-plots obtained with the four fitted models applied to different simulated scenarios are presented. In all cases where stratification was simulated, models without polygenic effect showed excess of small *P*-values, particularly with haplotypes which tend to capture more

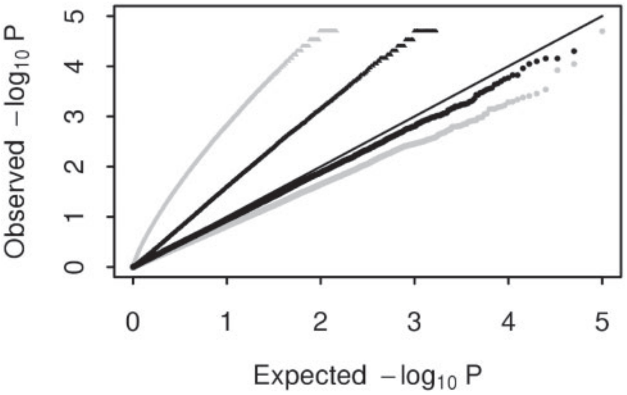

Fig. 1. QQ-plots obtained with marker SNPs (black) or ancestral haplotypes (gray) with (circles) or without (triangles) polygenic terms accounting for stratification included in the model

Table 1. Regression coefficients of QQ-plots obtained with four fitted models on four simulated designs [*P*-values computed with permutation test or with a gamma approximation test (in parenthesis)]

Simulated breed effect	Simulated polygenic variance	Fitted model			
		without polygenic effect		with polygenic effect	
		SNP[a]	Anc. Hap.[a]	SNP	Anc. Hap.
0	0	1.03 (1.01)	0.99 (1.02)	0.97 (0.96)	0.86 (0.87)
0.4	0	1.59 (1.57)	2.54 (3.06)	0.94 (0.94)	0.82 (0.83)
0	0.16	1.39 (1.36)	2.02 (2.22)	0.95 (0.95)	0.86 (0.86)
0.2	0.16	1.65 (1.63)	3.29 (3.89)	0.94 (0.94)	0.82 (0.83)

[a] Association is performed either with SNP or ancestral haplotypes.

stratification effects. The inclusion of a polygenic effect resulted in regression coefficients below 1.0, even when both breed and polygenic effects were simulated. In these simulations, two methods were used to estimate *P*-values, namely the permutation test and the gamma approximation test, with both yielding approximately the same regression slopes after correction for stratification.

Table 2 compares the power of models with marker SNPs or haplotypes for different OR and frequencies of the major variant (permutations were used to estimate *P*-values). In GWAS, due to multiple testing, low *P*-values must be achieved but the number of simulations allowed us only to estimate the power at $\alpha = 0.001$.The marker with the strongest association is not always the closest to the major variant. Therefore, power was tested in a 2 Mb window centered on the major variant, spanning ~30 SNPs. To correct for the resulting multiple testing (and correlation among successive tests along the region), chromosomes were randomly shuffled across individuals 10 000 times. For each permutation, the best *P*-value was stored and the association test was declared significant at $P < 0.001$ if one of the *P*-values was lower than 9990 of the best *P*-values obtained by permutation. Major variants, with resulting OR equal to 1.65 and 2.22, accounted only for a small fraction of the variation (r^2 between the SNP and the binary trait below 0.04 and 0.08 according to the SNP effect). Due to incomplete LD with causative SNPs, haplotypes and genotyped SNPs captured an even smaller fraction of that variance. Therefore the power was low (tests are already corrected for ~30 repeated correlated tests by the permutation procedure explained above), particularly when minor allele frequency (MAF) was below 0.2 and for small OR (1.65). For larger SNP effects, power increased, particularly when using ancestral haplotypes in the model which proved better in the present simulations for MAF above 0.10.

The power was also compared for different levels and types of structures (Table 3). In all cases, use of ancestral haplotypes resulted in higher power than for single SNPs and power decreased in datasets presenting structure (particularly for haplotypes and in presence of polygenes).

Table 2. Variant detection power ($\alpha = 0.001$) with SNP or ancestral haplotypes in a design without structure (no breed and polygenic effects) for different minor allelic frequency (MAF) classes of the causal SNP

MAF Class	OR of variant = 1.65			OR of variant = 2.22		
	r^{2a}	SNP	Anc. Hap	r^2	SNP	Anc. Hap
0.00–0.10	0.008	0.003	0.002	0.018	0.020	0.016
0.10–0.20	0.017	0.013	0.030	0.040	0.072	0.194
0.20–0.30	0.024	0.033	0.071	0.057	0.129	0.421
0.30–0.40	0.030	0.050	0.113	0.071	0.136	0.520
0.40–0.50	0.032	0.053	0.152	0.076	0.150	0.622

[a] r^2 between causative variant and observed phenotype.

Table 3. Power of association mapping ($\alpha = 0.001$) with SNP and haplotypes in different designs with stratification (statistics are provided across all MAFs)

Breed effect	Polygenic variance	OR of variant = 1.65		OR of variant = 2.22	
		SNP	Anc. Hap.	SNP	Anc. Hap.
0	0	0.023	0.078	0.079	0.373
0	0.16	0.025	0.057	0.077	0.322
0.4	0	0.027	0.067	0.077	0.359
0.7	0	0.028	0.071	0.075	0.365
0.2	0.16	0.023	0.053	0.073	0.322

4.2 Real data

Data from the Belgian Blue cattle breed heredo-surveillance platform were used to test the method on real datasets. Four phenotypes were analyzed: 3 monogenic recessive diseases [gingival hamartoma (33 cases), arthrogryposis (13 cases) and prolonged gestation (25 cases)] and color-sidedness which is monogenic dominant (8 cases). The causative variants are known for hamartoma (Sartelet *et al.*, in preparation) and color-sideness (Durkin *et al.*, 2012) whereas for arthrogryposis and prolonged gestation, diagnostic tests have been developed based on markers in LD with the causative variants. In addition to cases, genotypes from 300 controls were available. Individuals were genotyped for a custom made 50 K bovine chip described in Charlier *et al.* (2008). After deleting markers with a call rate below 0.90 or having a MAF below 0.05, 41 878 SNPs mapping to autosomal chromosomes were used in the study. Haplotypes were reconstructed using DualPHASE (Druet and Georges, 2010) with 10 ancestral haplotypes (we reduced the number of ancestral haplotypes to 10 since the number of individuals is much smaller than in the simulation study). Association studies were also performed with EMMAX (Kang *et al.*, 2010) which performs single point (SNPs) association studies with LMM that account for stratification (through inclusion of a kinship matrix). After Bonferroni correction for ~50 000 tests, genome-wide significance was set at 10^{-6}.

For the 3 monogenic recessive diseases *P*-values (estimated with the gamma approximation) below 10^{-40} were obtained (Manhattan plots are available in Supplementary Figures) in regions in which almost all cases were homozygous for a specific ancestral haplotype whereas almost none of the controls was homozygous for that haplotype, suggesting high LD between this ancestral haplotype and the causative variant. The identified regions were in agreement with the previous findings.

The Manhattan plot for color-sidedness is presented in Figure 2a. The lowest *P*-value is below 10^{-10} and the corresponding position is located at 0.7 Mb from a CNV causing the phenotype, a copy of a chromosomal segment on BTA6 encompassing the KIT locus which translocated to BTA29 (Durkin *et al.*, 2012). All color-sided individuals carry at least one copy of the same ancestral haplotype (it has dominant behavior). Some controls also carry the haplotype but the phenotype is not observed. Indeed, the phenotype has an incomplete penetrance since it cannot be observed on individuals with completely white coats (homozygous genotypes for a frequent common codominant mutation at the roan locus results in white coats [Charlier *et al.*, 1996)].

Associations performed with EMMAX are presented in Supplementary Figures for monogenic recessive diseases and Figure 2b for color-sideness. This software relies on LMM and assumes that the traits are normally distributed. As other LMM packages, it can still be applied on binary traits and perform well as shown in Supplementary Figures where several SNP in the region surrounding the causative SNP were highly significant and no other SNP reached such levels of significance.

However, some SNPs reach genome-wide significance in non-causative regions (e.g. association study for arthrogryposis). For color-sideness, use of EMMAX resulted in a Manhattan plot where the causative region is non-significant and difficult to identify (other regions of the genome show higher significance). These examples illustrate that considering binary trait as normally distributed and using SNP as covariates can result in situations where the region harboring the causative mutation is difficult to identify. In such situations, extension to GLMM and use of ancestral haplotypes resulted in associations where many positions in the region of interest have high level of significance, clearly above the remainder of the genome and with less non-causative regions reaching genome-wide significance.

4.3 Computational performance

To test the computational efficiency of the developped software, we ran an implementation compiled with Intel Fortran using openMP and MKL libraries on a dataset with a total of 4600 individuals genotyped for 500 000 SNPs. The analysis was performed on a Intel Xeon E5520 processor at 2.27 GHz using four threads. The full analysis lasted 2 h 36 min and required a total of 2.28 Gb of memory.

5 DISCUSSION

Our simulation studies showed that taking into account genomic relationships among individuals through inclusion of a polygenic effect in GLMM accounted for stratification, as previously observed with LMM (e.g. Malosetti *et al.*, 2007; Yu *et al.*, 2006; Zhao *et al.*, 2007). It was also observed that corrections were effective in structured populations (breed effect) and/or when family structure was present (polygenic effect). Most designs in model organisms, plants or animal species present a high level of stratification because either several populations are used in the study or the individuals are closely related, making these robust corrections essential. In addition to correcting for stratification, the GLMM framework offers additional flexibility because it allows for a better modeling of the phenotypes through the inclusion of additional covariates (e.g. sex, age, etc.) and a consequently better association study where all known nuisance factors have been corrected for. Such a possible nuisance factor could be a measure of the population structure, as suggested in some studies (e.g. Price *et al.*, 2006; Pritchard *et al.*, 2000). The effect of adding this correction concurrently with the genomic relatedness structure might be necessary in some populations albeit not always (Kang *et al.*, 2008; Zhao *et al.*, 2007). Another point stressing flexibility of LMM is that association can easily be performed with either SNPs or (ancestral) haplotypes. Extension from LMM to GLMM is important for traits that are not normally distributed. However, in many situations such as large balanced case/control studies where variants are not very rare and have low or moderate effects, LMM perform well with binary traits. GLMM are recommended for strong deviations from normality, when cases or controls are rare within a cell (covariates of the model such as fixed, SNP or haplotype effects). Such situations occur more often in smaller designs with few cases (or controls) and with rare variants (or haplotypes). Such designs are still common in animal or plant species and our applications to real data illustrate that in such cases, use of a GLMM with a logit link function results in cleaner association than LMM.

Slopes of QQ-plots indicated that statistical tests were too conservative resulting in a loss of power. Similar deflation has been described in Amin *et al.* (2007) and is due to the fact that polygenic effects (used for correction) and SNP or haplotype effects are correlated (e.g. SNP or haplotypes are used to estimate *K*). This correlation is stronger with haplotype which are therefore more affected by the over-correction. One solution would be to correct for deflation with genomic control as in Amin *et al.* (2007) or to perform a test modeling simultaneously polygenic and haplotype effects at most interesting positions. Without such corrections, the proposed model may be more conservative than some other methods relying on LMM.

Although computational efficiency was not the main goal of the present study, the developed method presents interesting computational features. First, GLMM must be solved only once (with the SNP or haplotype variance set to 0). Solving the GLMM equations and inferring the variance components

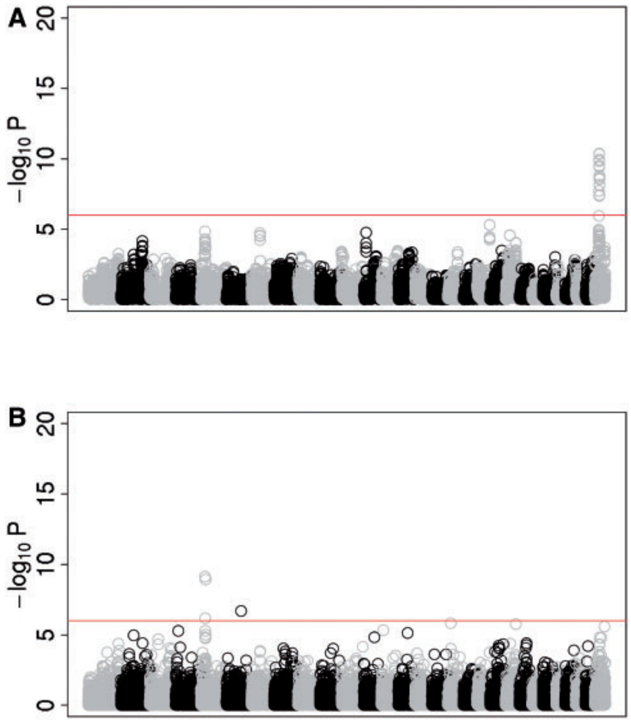

Fig. 2. Manhattan plot for association study for color-sideness with (**A**) GLMM score tests using ancestral haplotypes and (**B**) EMMAX

are potentially time consuming. With likelihood ratio tests, variances are inferred for a model including polygenic and SNP (or haplotype) random effects for each tested position. With the present method, the test score is obtained from a simple statistic based on records corrected for effects of the model under the null hypothesis (without SNP or haplotypic effect). Performing this test is much faster than inferring variance in the full model. This is very similar to the approach used in GRAMMAR (Aulchenko *et al.*, 2007). In addition, since the data are corrected for family and population structure it becomes exchangeable and permutations can be performed freely, which was not the case for the raw data. Thanks to permutations, empirical *P*-values corrected for multiple testing can easily be obtained. Still, with permutations it would be time consuming to obtain small *P*-values as typically needed in GWAS studies; in that case, approximation of the test score distribution with a gamma distribution (with scale and shift parameters obtained empirically through 1000 permutations) seems to perform well.

As in previous studies (Druet and Georges, 2010; Su *et al.*, 2008), assigning chromosomes to ancestral haplotypes resulted in high LD between haplotype groups and underlying mutations. In the present study, the LD was much higher than when using SNP for a cattle population and with ~50 000 SNPs covering the genome. The picture might change with different marker densities or in other populations. The method for clustering haplotypes has also been successfully applied to the fine-mapping of a QTL affecting bovine stature (Karim *et al.*, 2011) in a crossbred population. In that study, the association between the ancestral haplotypes and the later candidate causative variants was almost perfect (2 misclassified haplotypes out of 1490). More recently, our method allowed to fine-map a mutation causing dwarfism in cattle which was always associated to the same ancestral haplotype (Sartelet *et al.*, 2012). Other examples in the present study illustrate the high LD between the ancestral haplotypes obtained with DualPHASE (Druet and Georges, 2010) and ungenotyped variants. The association should be better for more recent variants which rapidly increased in frequency due to selection. In that case, the length of the haplotype associated to the variant would be longer than for random variants (as those used in the simulation study), making it easier to identify the haplotype. Ancestral haplotypes can be associated with different types of variants including SNP, multiple alleles, several-linked SNP (a small haplotype), insertions/deletions and duplications. In the real data example on color-sidedness, ancestral haplotypes presented high LD with a trans CNV and other examples of association between ancestral haplotypes and deletions or duplications (either in cis or trans position) can be found in Durkin *et al.* (2012). Note, when large reference populations genotyped at high density (or sequenced) are available, imputation followed by single point association would probably result in higher power than use of ancestral haplotypes (if there is only one causal variant and if SNPs in high LD with this variant are genotyped in the reference panel). However, such reference populations are only available in a few species.

While providing high LD with underlying variants, the use of ancestral haplotypes also controls the number of haplotype groups to be used in the study, which is important to maintain statistical power. This method proves also flexible since there is no need to define arbitrary windows and since haplotype origin can change at any position along the chromosome. For instance, recombinant haplotypes do not create additional haplotypes groups: they are simply potentially assigned to different groups on each side of the cross-over position. Finally, haplotypes with small differences due to genotyping errors or new non-causative mutations can still be grouped together whereas with less flexible methods, new haplotype groups would be defined for each difference, resulting in a loss of power. For the same reason, missing genotypes are easily handled. Our score test framework can easily be applied with other methods for clustering haplotypes, even those modeling a correlation between haplotypic effects. Our method does not rely on a particular biological model but identifies ancestral haplotypes significantly associated with the disease. Therefore it can be applied to monogenic recessive diseases, dominant diseases, phenotypes with complete or incomplete penetrance, oligenic or polygenic diseases or complex traits. It is also robust to misclassified samples which will only reduce slightly the power since there are no strong assumptions such as sharing of an IBD segment in all cases. For instance, the method was used to fine-map a variant causing dwarfism in Belgian Blue cattle (Sartelet *et al.*, 2012) which was the cause only for a subset of cases (14 out of 33). Still, the method detected with high significance (*P*-value $<10^{-11}$) the region harboring the causative mutation. Further veterinary examination revealed that dwarfs could be classified into different categories and that the 14 cases corresponded to a specific sub-group. Even without that knowledge, the variant was identified, stressing the robustness of the method. The example of color-sidedness also illustrates that with only a few cases (8), a dominant gene (cases carry only one haplotype) and incomplete penetrance (the phenotype is not observed on white animals) the method still identifies with high significance the region harboring the causative variant.

ACKNOWLEDGEMENTS

We acknowledge LIC for providing data. Tom Druet and Carole Charlier are, respectively, Research Associate and Senior Research Associate of the Fonds de la Recherche Scientifique—FNRS. The authors thank the GIGA Bioinformatics platform and SEGI (University of Liège) for providing computing resources.

Funding: This work was funded by grants from the Walloon Ministry of Agriculture (Rilouke) and the European Union (PIGENDEF).

Conflict of Interest: None declared.

REFERENCES

Amin,N. *et al.* (2007) A genomic background based method for association analysis in related individuals. *PLoS One*, **2**, e1274.

Aranzana,M.J. *et al.* (2005) Genome-wide association mapping in arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS Genet.*, **1**, e60.

Aulchenko,Y.S. *et al.* (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, **177**, 577–585.

Blott,S. *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.

Breslow,N. and Clayton,D. (1993) Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.*, **88**, 9–25.

Browning,S.R. (2008) Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum. Genet.*, **124**, 439–450.

Browning,S.R. and Browning,B.L. (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. *et al.* (1996) Microsatellite mapping of the bovine roan locus: a major determinant of white heifer disease. *Mamm. Genome*, **7**, 138–142.

Charlier,C. *et al.* (2008) Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nat. Genet.*, **40**, 449–454.

de Roos,A.P.W. *et al.* (2011) Genomic breeding value estimation using genetic markers, inferred ancestral haplotypes, and the genomic relationship matrix. *J. Dairy Sci.*, **94**, 4708–4714.

Devlin,B. and Roeder,K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Donnelly,P. (2008) Progress and challenges in genome-wide association studies in humans. *Nature*, **456**, 728–731.

Druet,T. and Georges,M. (2010) A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics*, **184**, 789–798.

Druet,T. *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.

Durkin,K. *et al.* (2012) Serial translocation via circular intermediates underlies color-sidedness in cattle. *Nature*, **482**, 81–84.

Durrant,C. *et al.* (2004) Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am. J. Hum. Genet.*, **75**, 35–43.

Eding,H. *et al.* (2001) Marker-based estimates of between and within population kinships for the conservation of genetic diversity. *J. Anim. Breed. Genet.*, **118**, 141–159.

George,A.W. *et al.* (2000) Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics*, **156**, 2081–2092.

Hayes,B.J. and Goddard,M.E. (2008) Technical note: prediction of breeding values using marker-derived relationship matrices. *J. Anim. Sci.*, **86**, 2089–2092.

Kang,H.M. *et al.* (2008) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709–1723.

Kang,H.M. *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–354.

Karim,L. *et al.* (2011) Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. *Nat. Genet.*, **43**, 405–413.

Malosetti,M. *et al.* (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to phytophthora infestans in potato. *Genetics*, **175**, 879–889.

McCullagh,P. and Nelder,J. (1989) *Generalized Linear Models (Monographs on Statistics and Applied Probability 37),* Chapman Hall, London.

Meuwissen,T.H. and Goddard,M.E. (2001) Prediction of identity by descent probabilities from marker-haplotypes. *Genet. Sel. Evol.*, **33**, 605–634.

Price,A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

Pritchard,J.K. *et al.* (2000) Association mapping in structured populations. *Am. J. Hum. Genet.*, **67**, 170–181.

Sartelet,A. *et al.* (2012) A splice site variant in the bovine RNF11 gene compromises growth and regulation of the inflammatory response. *PLoS Genet.*, **8**, e1002581.

Schaid,D.J. *et al.* (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.*, **70**, 425–434.

Scheet,P. and Stephens,M. (2006) A fast and flexible statistical model for largescale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, **78**, 629–644.

Seltman,H. *et al.* (2003) Evolutionary-based association analysis using haplotype data. *Genet. Epidemiol.*, **25**, 48–58.

Su,S.-Y. *et al.* (2008) Disease association tests by inferring ancestral haplotypes using a hidden Markov model. *Bioinformatics*, **24**, 972–978.

Threadgill,D.W. *et al.* (2002) Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mamm. Genome*, **13**, 175–178.

Tzeng,J.-Y. and Zhang,D. (2007) Haplotype-based association analysis via variance components score test. *Am. J. Hum. Genet.*, **81**, 927–938.

Verbeke,G. and Molenberghs,G. (2003) The use of score tests for inference on variance components. *Biometrics*, **59**, 254–262.

Yu,J. *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, **38**, 203–208.

Zhao,K. *et al.* (2007) An Arabidopsis example of association mapping in structured samples. *PLoS Genet.*, **3**, e4.